COS 513: SEQUENCE MODELS I

LECTURE ON NOV 22, 2010

PREM GOPALAN

1. INTRODUCTION

In this lecture we consider how to model sequential data. Rather than assuming that the data are all independent of each other we assume they come in sequence $X_{1..T} = x_1, x_2.., x_T$. There are two types of sequential models that are quite similar to each other: *Hidden Markov Model (HMM)* and *Kalman Filter*. This lecture focuses on HMM which has many applications including genome modeling and action recognition.

HMMs are a generalization of the finite mixture model (MM) to sequences. In MM, the process of generating IID data involves choosing a

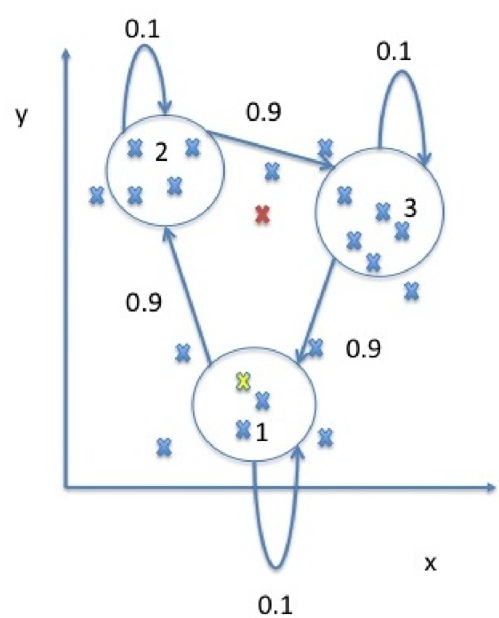

FIGURE 1. *Diagram representing transitions between mixture components 1, 2, 3 and observed data. The probability of transition is shown on the edges. xs represent data points and x_1 and x_2 are indicated by the yellow and red x respectively.*

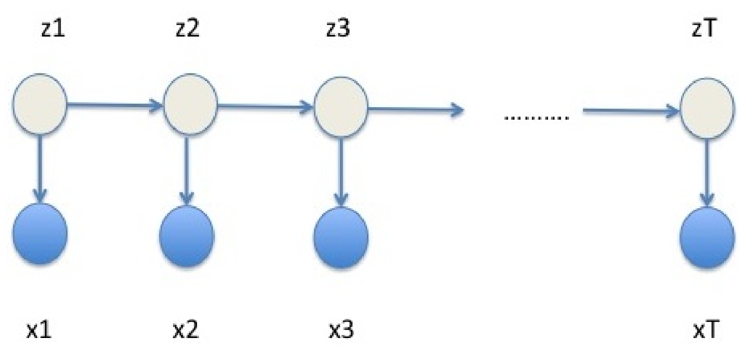

FIGURE 2. Graphical Model

component according to a distribution $p(z)$, independent of choice of components in other steps, and choosing a data vector from the distribution, $p(x|z)$. In HMM, the mixture component is chosen dependent on the previous component. Each component can be seen as a state, and we augment the basic MM to include a matrix of *transition probabilities*.

Figure 1 illustrates this difference. The xs are elements of the sequence. Let the *yellow* x represent x_1 and the *red* x represent x_2. Then in MM, x_2 is approximately equally likely to belong to component 2 or 3. In HMM, x_2 is more likely to belong to component 2, since x_1 belongs to 1, and the probability of state transition from 1 to 2 is high.

2. GRAPHICAL MODEL FOR HMM

In Figure 2, each of the z_t is a multinomial random variable represented by a indicator vector of size K, whose component i is 1 if the cluster index i (for the clusters associated with data $x_{1:T}$) is indicated, and 0 if not. For a particular configuration $(z, y) = (z_1, z_2, .., z_T, x_1, x_2, ..x_T)$ as shown in Figure 2, the joint probability is given by the product of local conditional probabilities as follows:

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

We assume above that the distribution $p(x_t|z_t)$ is independent of t.

2.1. Emission probabilities.

For a given state k, there is a set of emission probabilities governing the distribution of y_t and we represent it by θ_k. For example, θ_k could be a parameter to a multivariate Gaussian or multinomial Poisson. Thus $p(x_t|z_t)$ can be written as:

$$p(x_t|z_t) = \prod_{k=1}^{K} p(x_t|\theta_k)^{z_t^k} \tag{2}$$

2.2. **Transition probabilities.** Define a K x K state transition matrix A, where each entry a_{ij} is the probability $p(z_t^j = 1|z_{t-1}^i = 1)$. The probability of the next state z_t given the current z_{t-1} is given by:

$$p(z_t|z_{t-1}) = \prod_{k=1}^{K} \prod_{j=1}^{K} [a_{jk}]^{z_{t-1}^j z_t^k} \tag{3}$$

Since only one component of z_t or z_{t-1} is 1, there is only one factor on the right-hand side that is different from one.

2.3. **Initial distribution.** The first state node in the sequence has no parents. Thus we define π to be the distribution where $\pi_k = p(z_1^k = 1)$. A more formal definition is as follows:

$$p(z_1) = \prod_{k=1}^{K} \pi_k^{z_1^k} \tag{4}$$

2.4. **Conditional independence.** From the graphical model, and using Bayes ball, we can see that conditioning on z_{t-1} renders z_t and z_{t-2} independent. Thus the future is independent of the past, given the present. This is the *Markov property*. Note that this is not true when conditioned on the output node x_{t-1} instead of z_{t-1}.

3. ESTIMATING HMM PARAMETERS USING THE EM ALGORITHM

The parameters of an HMM include the emission probabilities $\hat{\theta}$, the transition matrix $\hat{A}$ and the initial probability distribution $\hat{\pi}$. Given data $x_{1..T}$, we want to estimate these parameters. First we write down the expected complete log likelihood using equations 1 to 4 with respect to the posterior $p(z_{1..T}|x_{1..T})$:

(5)

$$\mathbb{E}[logp(x_{1..T}, z_{1..T})] = \mathbb{E}[log\{\prod_{k=1}^{K} \pi_k^{z_1^k} \prod_{t=2}^{T} \prod_{j=1}^{K} \prod_{k=1}^{K} [a_{jk}]^{z_{t-1}^j z_t^k} \prod_{t=1}^{T} \prod_{k=1}^{K} p(x_t|\theta_k)^{z_t^k}\}]$$

(6)
$$= \sum_{k=1}^{K} \mathbb{E}[z_1^k]\ log\pi_k + \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}[z_{t-1}^j z_t^k]\ log[a_{jk}] + \sum_{t=1}^{T}\sum_{k=1}^{K} \mathbb{E}[z_t^k]\ logp(x_t|\theta_k)$$

E step. We need to compute the following conditional expectations. We will return to these expectations at the end of this section.

$$\mathbb{E}[z_t^k] = p(z_t = k|x_{1..T}) \tag{7}$$

$$\mathbb{E}[z_{t-1}^j z_t^k] = p(z_{t-1} = j, z_t = k|x_{1..T}) \tag{8}$$

M step. In the M step, the parameters are adjusted using a process that is equivalent to assuming that the latent variables have been observed. Holding the above expectations fixed, we optimize the parameters to try to eventually converge to a maximum likelihood estimate. An estimate for the prior probability of state z_1, π_k is given by:

$$\pi_k = \mathbb{E}[z_1^k] / \sum_{j=1}^{k} \mathbb{E}[z_1^j] \tag{9}$$

We then estimate the probability of moving from j^{th} state to k^{th} state. In equation 10, the numerator is the number of transitions from j^{th} to k^{th} state and the denominator the total number of transitions from j^{th} state.

$$a_{jk} = \sum_{t=2}^{T} \mathbb{E}[z_{t-1}^j z_t^k] / \sum_{t=2}^{T}\sum_{l=1}^{k} \mathbb{E}[z_{t-1}^j z_t^l] \tag{10}$$

θ_k is estimated as the weighted maximum likelihood estimate with weights given by $\mathbb{E}[z_t^k]$. For example, in the Gaussian case, μ_k, the cluster center, is estimated as follows:

$$\mu_k = \sum_{t=1}^{T} \mathbb{E}[z_t^k] x_t / \sum_{t=1}^{T} \mathbb{E}[z_t^k] \tag{11}$$

Each term in the numerator in equation 11 is the probability of x_t being in cluster k multiplied by x_t, and the denominator is the expected number of data points in cluster k. The multinomial case where each x_t has exactly one of D fixed, finite outcomes, is as follows:

$$p(x_t|\theta_k) = \prod_{i=1}^{D} \theta_{k,i}^{x_t^i} \tag{12}$$

$$\theta_{k,i} = \sum_{t=1}^{T} \mathbb{E}[z_t^k] x_t^i / \sum_{t=1}^{T} \mathbb{E}[z_t^k] \tag{13}$$

Now, let us consider how to compute $\mathbb{E}(z_t|x_{1..T})$ and $\mathbb{E}[z_{t-1}, z_t|x_{1..T}]$ in the E step. Define $\alpha(z_t)$, $\beta(z_t)$ as follows using a simple application of the Bayes rule, chain rule and conditional independence.

$$\begin{aligned}
\mathbb{E}[z_t|x_{1..T}] &= p(z_t|x_{1..T}) \\
&= p(z_t, x_{1..T})/p(x_{1..T}) \\
&= p(x_{1..t}, z_t).p(x_{t+1..T}|z_t)/p(x_{1..T}) \\
&= \alpha(z_t).\beta(z_t)/p(x_{1..T})
\end{aligned}$$

$\alpha(z_t)$ is the probability of emitting a sequence of outputs $x_{1..t}$ and ending up in state z_t. $\beta(z_t)$ is the probability of emitting a sequence of outputs $x_{t+1..T}$ starting from state z_t.

$$\begin{aligned}
\mathbb{E}[z_{t-1}, z_t|x_{1..T}] &= p(z_{t-1}, z_t|x_{1..T}) \\
&= p(x_{1..T}, z_{t-1}, z_t)/p(x_{1..T}) \\
&= p(x_{1..t-1}, z_{t-1}).p(x_{t..T}, z_t|x_{1..t-1}, z_{t-1})/p(x_{1..T}) \\
&= p(x_{1..t-1}, z_{t-1}).p(z_t|z_{t-1}).p(x_{t..T}|z_t, z_{t-1})/p(x_{1..T}) \\
&= p(x_{1..t-1}, z_{t-1}).p(z_t|z_{t-1}).p(x_t|z_t, z_{t-1}).p(x_{t+1..T}|z_t, z_{t-1})/p(x_{1..T}) \\
&= p(x_{1..t-1}, z_{t-1}).p(z_t|z_{t-1}).p(x_t|z_t).p(x_{t+1..T}|z_t)/p(x_{1..T}) \\
&= \alpha(z_{t-1}).p(z_t|z_{t-1}).p(x_t|z_t).\beta(z_t)/p(x_{1..T})
\end{aligned} \tag{14}$$

In the above sequence of equations, step 3 follows from splitting the sequence $x_{1..T}$ into $x_{1..t-1}$ and $x_{t..T}$, and applying Bayes rule. In step 4, we use the independence of z_t from $x_{1..t-1}$ given z_{t-1}, and the independence of $x_{t..T}$ from $x_{1..t-1}$ given z_t. Steps 5 and 6 use the independence of x_t from z_{t-1} and $x_{t+1..T}$ from z_{t-1}, and from each other, given z_t. Note that $p(z_t|z_{t-1})$ is given by $a_{z_t,z_{t-1}}$. In the next lecture, we will consider algorithms to compute $\alpha(z_t)$ and $\beta(z_t)$.