STA561: Probabilistic Machine Learning

Lecture - Hidden Markov Models (9/23/13)

Lecturer: Barbara Engelhardt *Scribe names: Yue Jiang, Yi Yin, Xiujin Guo, Shuyang Yao*

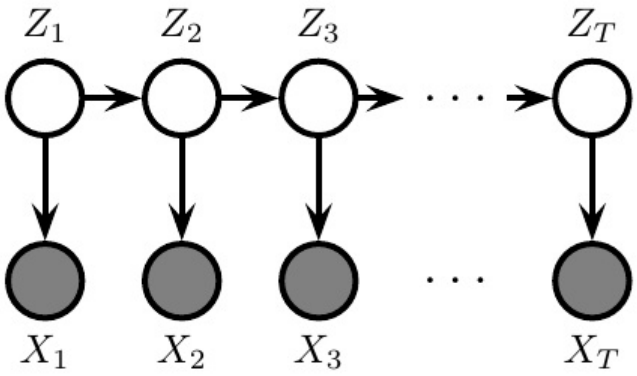
1 Hidden Markov Models

Hidden Markov models are widely used to model potentially complex processes which take place over time. Common examples include analyzing trends in the stock market, automatic speaker recognition, gesture recognition, gene finding, and as a building block for weather prediction Spatiotemporal models.

The fundamental idea behind *hidden Markov models* is that you may be able to express your complicated observed data, $X \in \mathbb{R}^d$, in terms of some *hidden* (unobserved) data, Z, which have a simple Markov structure. The (first order) *markov property* states that the future is conditionally independent of the past given the present.

$$(Z_{t+1:T} \perp Z_{1:t-1}) \mid Z_t$$

For now, we will suppose that the hidden variables in the model have *discrete states*, e.g. $Z_t \in \{1, ..., k\}$. We will also suppose that our model is *discrete time*. That is to say, we only care about the state of the random process at some discrete set of time points, e.g. $\{1, 2, \ldots T\}$.



We can easily write out the joint distribution of the observed and latent data using the graph above.

$$\mathsf{p}(Z_1, ..., Z_T, X_1, ...X_T) = \mathsf{p}(Z_1) \left[\prod_{t=2}^{T} \mathsf{p}(Z_t \mid Z_{t-1})\right] \left[\prod_{t=1}^{T} \mathsf{p}(X_t|Z_t)\right]$$

The *emission probabilities* determine the distribution of the observed data X_t given the hidden data Z_t. That is, the *emission probability* at time t is given by $\mathsf{p}(X_t = x_t \mid Z_t = k, \theta) = \eta$. For now, suppose that X_t is multivariate normal given Z_t so that $\eta = \mathcal{N}(x_t|\mu_k, \Sigma_k)$.

Let's write $z_t = [0, 1, 0]^T$ as a multinomial vector (in this example, K=3, $Z_t = 2$). The *transition probability* at time t gives the probability of the next latent state given the current latent state. It is convenient to

collect these into a transition matrix, A, where $A_{kj} = \mathsf{p}(z_t^j \mid z_{t-1}^k)$. That is, the k, j^{th} entry of A gives the probability of transitioning from state k to state j. Note that transition matrix is homogeneous over time. This implies that the underlying Markov chain is at its stationary distribution.

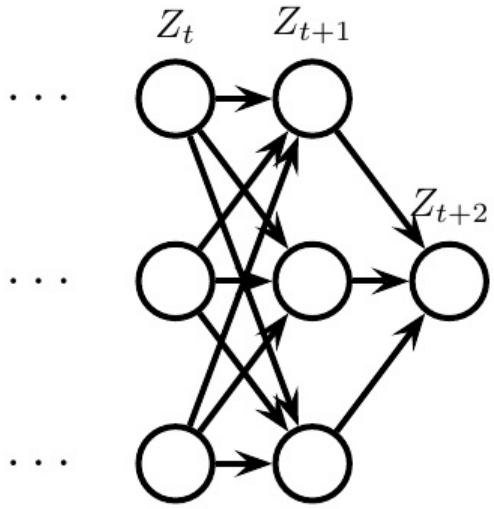
2 Types of Inference

1. "Filtering": compute a belief state $\mathsf{p}(Z_t|X_{1:t})$.
   As we collect online data, all previous data $X_{1:t}$ are used to estimate Z_t. This can not be simplified because we are not conditioning on Z_{t-1}

2. "Smoothing": compute $\mathsf{p}(Z_t|X_{1:T})$.
   This method utilizes all of the future data, $X_{t+1:T}$ in addition to the data collected up to time t to determine the distribution of the current latent state, Z_t.

3. Prediction: predict the future given the past: $\mathsf{p}(Z_{t+h}|X_{1:t})$. For example, let's suppose that $h = 2$:

$$\mathsf{p}(Z_{t+2} = z_{t+2}|X_{1:t}) = \sum_{z_{t+1}} \sum_{z_t} \underbrace{\mathsf{p}(z_{t+2} \mid z_{t+1})\mathsf{p}(z_{t+1} \mid z_t)\mathsf{p}(z_t \mid X_{1:t})}_{\text{power up transition matrix}}$$



4. MAP estimation / "Viterbi decoding": $\arg\max_{z_{1:T}} \mathsf{p}(Z_{1:T} = z_{1:T} \mid X_{1:T})$.
   The idea here is to obtain the most probable latent state sequence given the observed data.

5. Posterior sampling - $Z_{1:T} \sim \mathsf{p}(Z_{1:T} \mid X_{1:T})$
   Sampling can be useful for identifying where there is uncertainty in your latent variable estimation 'path'.

6. Probability of evidence - $\mathsf{p}(X_{1:T}) = \sum_{z_{1:T}} \mathsf{p}(X_{1:T}, z_{1:T}|\theta)$
   Note that we are summing over all possible sequences of hidden data. This is useful to obtain the probability of the data for the purposes of anomaly detection.

3 EM Algorithm

Model Parameters: $\theta = (A, \pi, \eta)$
The parameter A denotes the (stationary) transition matrix. Note that the rows of A must sum to 1. The parameter π specifies the initial latent state distribution, and the η parameters give the emission probabilities.

Observed data: $\mathcal{D} = \{(X_{1:T})_1, ..., (X_{1:T})_n\}$

Let's derive the EM algorithm:

1. Write out the complete log likelihood (n=1)

$$\begin{aligned}\ell_c(\theta, Z, X; \mathcal{D}) &= \log[\mathsf{p}(Z, X \mid \theta)] \\ &= \log\left\{\mathsf{p}(Z_1)\left[\prod_{t=1}^{T}\mathsf{p}(Z_t \mid Z_{t-1})\right]\left[\prod_{t=1}^{T}\mathsf{p}(X_t \mid Z_t)\right]\right\} \\ &= \log \pi_{Z_1} + \sum_{t=1}^{T-1}\log a_{Z_t, Z_{t+1}} + \sum_{t=1}^{T}\log \mathsf{p}(X_t \mid Z_t)\end{aligned}$$

2. Write out the expected complete log likelihood

$$\begin{aligned}\mathbb{E}\left[\ell_c(\theta; \mathcal{D})\right] &= \mathbb{E}\left[\sum_{k=1}^{K} Z_1^k \log \pi_k + \sum_{t=1}^{T-1}\sum_{j,k=1}^{K} Z_t^j Z_{t+1}^k \log a_{j,k} + \sum_{t=1}^{T}\log \mathsf{p}(X_t|Z_t, \eta)\right] \\ &= \sum_{k=1}^{K}\mathbb{E}[Z_1^k]\log \pi_k + \sum_{t=1}^{T-1}\sum_{j,k=1}^{K}\mathbb{E}[Z_t^j Z_{t+1}^k]\log a_{jk} + \sum_{t=1}^{T}\mathbb{E}[\log \mathsf{p}(X_t|Z_t, \eta)]\end{aligned}$$

3. Our expected sufficient statistics:
   E-step:

$$\mathbb{E}[Z_1^k] = \mathbb{E}[Z_1^k \mid X_{1:T}, \theta] = \mathsf{p}(Z_1^j = 1 \mid X_{1:T}, \theta)$$

   This is what we expect since Z_1 follows a Multinomial distribution, so its expectation is simply the vector of posterior probabilities. Furthermore, it is important to note that this is the same as performing smoothing.

$$\mathbb{E}[Z_t^j, Z_{t+1}^k] = \mathbb{E}[Z_t^j, Z_{t+1}^k \mid X_{1:T}, \theta] = \sum_{t=1}^{T-1}\mathsf{p}(Z_t^j Z_{t+1}^k \mid X_{1:T}, \theta)$$

   Note that intuitively, $\mathbb{E}[Z_t^j, Z_{t+1}^k]$ counts how often we see transition pairs. We can use the *forward - backward* algorithm to obtain this.

There will be more on the forward-backward algorithm, and the M-step next lecture.