CS 340 Lec. 21: Hidden Markov Models

AD

April 2011

Modelling Dependent Data

- For the time being, we have always assumed that available data $\{\mathbf{x}_t\}_{t=1}^{T}$ are independent.
- In numerous applications, we only have access to data which are statistically dependent; i.e.

$$p\left(\{\mathbf{x}_t\}_{t=1}^{T}\right) \neq \prod_{t=1}^{T} p\left(\mathbf{x}_t\right).$$

- Typical applications include: speech processing, tracking, stock prices.
- Most popular model for time dependent data is Hidden Markov Models = Mixture Models + Markov chain on the "cluster labels".

Introduction to HMM

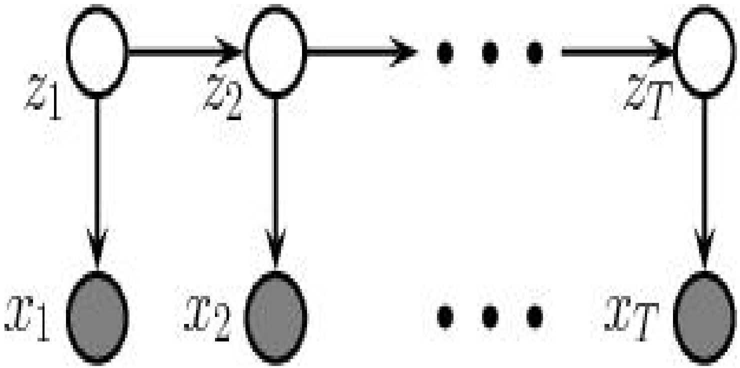
- In a standard mixture models, we have independent cluster labels $\{z_t\}_{t=1}^{T}$ and data $\{\mathbf{x}_t\}_{t=1}^{T}$ so

$$\begin{aligned} p\left(\{z_t\}_{t=1}^{T}, \{\mathbf{x}_t\}_{t=1}^{T}\right) &= \prod_{t=1}^{T} p\left(z_t, \mathbf{x}_t\right) = \prod_{t=1}^{T} p\left(z_t\right) p\left(\mathbf{x}_t \mid z_t\right) \\ &= \prod_{t=1}^{T} p\left(z_t\right) \prod_{t=1}^{T} p\left(\mathbf{x}_t \mid z_t\right) \end{aligned}$$

- In an HMM model, the cluster labels $\{z_t\}_{t=1}^{T}$ follow a Markov chain

$$p\left(\{z_t\}_{t=1}^{T}, \{\mathbf{x}_t\}_{t=1}^{T}\right) = p\left(z_1\right) \prod_{t=2}^{T} p\left(z_t \mid z_{t-1}\right) \prod_{t=1}^{T} p\left(\mathbf{x}_t \mid z_t\right)$$

Graphical Representation

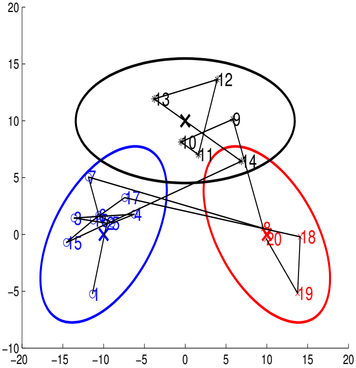


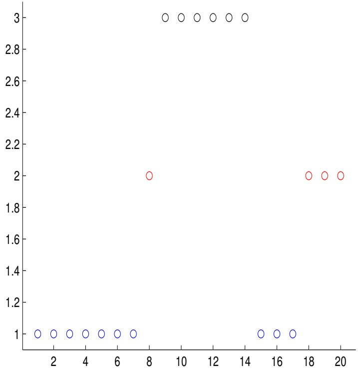
HMM as a directed graphical model

HMM Specification

- Assume $z_t \in \{1, ..., K\}$ then the Markov chain is defined by its initial distribution
$$p(z_1 = k) = \pi_k$$
and the transition probabilities
$$p(z_t = l | z_{t-1} = k) = P_{k,l}$$
- We have the conditional densities/distribution
$$p(\mathbf{x}_t | z_t = k) = p_k(\mathbf{x}_t)$$
where we could have for example $p_k(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \mu_k, \Sigma_k)$ or $p_k(\mathbf{x}_t = 1) = \alpha_k$.

Data Sampled from an HMM





(left) Some data sampled from a 3 state HMM. Each state emits from a 2d Gaussian. (right) The hidden state sequence $\{z_t\}$.

Applications of HMM

- *Automatic speech recognition*: $\mathbf{x}_t$ is the speech signal, z_t represents the word that is being spoken.
- *Activity recognition*: $\mathbf{x}_t$ represents a video frame, z_t is the class of activity the person is engaged in (e.g., running, walking, sitting, etc.)
- *Part of speech tagging*: $\mathbf{x}_t$ represents a word, z_t represents its part of speech (noun, verb, adjective, etc.)
- *Gene finding*: $\mathbf{x}_t$ represents the DNA nucleotides (A,C,G,T), z_t represents whether we are inside a gene-coding region or not.

Inference in HMM

- Assume for the time being that the parameters of the models are known, we want to estimate z_t given observations $\{\mathbf{x}_t\}$.
- **Filtering**: compute $p\left(z_t = k \middle| \mathbf{x}_{1:t}\right)$
- **Prediction:** compute $p\left(z_{t+L} = k \middle| \mathbf{x}_{1:t}\right)$ for $L > 0$
- **Smoothing**: compute $p\left(z_t = k \middle| \mathbf{x}_{1:T}\right)$
- In the independent case,
  $p\left(z_t = k \middle| \mathbf{x}_{1:t}\right) = p\left(z_t = k \middle| \mathbf{x}_{1:T}\right) = p\left(z_t = k \middle| \mathbf{x}_t\right)$ and
  $p\left(z_{t+L} = k \middle| \mathbf{x}_{1:t}\right) = p\left(z_{t+L} = k\right)$.

Inference in HMM: Filtering

- Given the filter $p\left(z_{t-1} \middle| \mathbf{x}_{1:t-1}\right)$ at time $t-1$, we compute $p\left(z_t \middle| \mathbf{x}_{1:t}\right)$ as follows.
- *Prediction*:

$$
\begin{aligned}
p\left(z_t = k \middle| \mathbf{x}_{1:t-1}\right) &= \textstyle\sum_{l=1}^{K} p\left(z_{t-1} = l, z_t = k \middle| \mathbf{x}_{1:t-1}\right) \\
&= \textstyle\sum_{l=1}^{K} p\left(z_t = k \middle| \mathbf{x}_{1:t-1}, z_{t-1} = l\right) p\left(z_{t-1} = l \middle| \mathbf{x}_{1:t-1}\right) \\
&= \textstyle\sum_{l=1}^{K} p\left(z_t = k \middle| z_{t-1} = l\right) p\left(z_{t-1} = l \middle| \mathbf{x}_{1:t-1}\right) \\
&= \textstyle\sum_{l=1}^{K} P_{l,k}\; p\left(z_{t-1} = l \middle| \mathbf{x}_{1:t-1}\right)
\end{aligned}
$$

- *Update:*

$$
\begin{aligned}
p\left(z_t = k \middle| \mathbf{x}_{1:t}\right) &= \frac{p\left(\mathbf{x}_t \middle| z_t = k\right) p\left(z_t = k \middle| \mathbf{x}_{1:t-1}\right)}{\sum_{l=1}^{K} p\left(\mathbf{x}_t \middle| z_t = l\right) p\left(z_t = l \middle| \mathbf{x}_{1:t-1}\right)} \\
&= \frac{p_k\left(\mathbf{x}_t\right) p\left(z_t = k \middle| \mathbf{x}_{1:t-1}\right)}{\sum_{l=1}^{K} p_l\left(\mathbf{x}_t\right) p\left(z_t = l \middle| \mathbf{x}_{1:t-1}\right)}
\end{aligned}
$$

- This has computational complexity $O\left(K^2 T\right)$.

Inference in HMM: Prediction

- We want to compute $p\left(z_{t+L}=k \middle| \mathbf{x}_{1:t}\right)$ for $L>0$.
- We have

$$p\left(z_{t+1}=k \middle| \mathbf{x}_{1:t}\right)=\sum_{l=1}^{K} P_{l,k}\ p\left(z_{t}=l \middle| \mathbf{x}_{1:t-1}\right)$$

and similarly

$$p\left(z_{t+m}=k \middle| \mathbf{x}_{1:t}\right)=\sum_{l=1}^{K} P_{l,k}\ p\left(z_{t+m-1}=l \middle| \mathbf{x}_{1:t-1}\right)$$

Inference in HMM: Smoothing

- We have for $1 \leq t < T$

$$p\left(z_t = k \middle| \mathbf{x}_{1:T}\right) = \frac{p\left(z_t = k \middle| \mathbf{x}_{1:t-1}\right) p\left(\mathbf{x}_{t:T} \middle| z_t = k\right)}{\sum_{l=1}^{K} p\left(z_t = l \middle| \mathbf{x}_{1:t-1}\right) p\left(\mathbf{x}_{t:T} \middle| z_t = l\right)}$$

- We can compute $p\left(\mathbf{x}_{t:T} \middle| z_t = k\right)$ using the following backward recursion initialized at $p\left(\mathbf{x}_T \middle| z_T = k\right) = p_k\left(\mathbf{x}_T\right)$

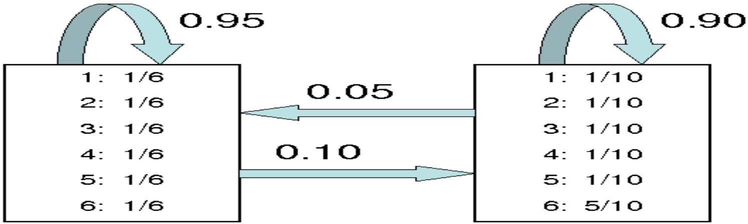
$$\begin{aligned} p\left(\mathbf{x}_{t+1:T} \middle| z_t = k\right) &= \textstyle\sum_{l=1}^{K} p\left(\mathbf{x}_{t+1:T}, z_{t+1} = l \middle| z_t = k\right) \\ &= \textstyle\sum_{l=1}^{K} p\left(\mathbf{x}_{t+1:T} \middle| z_t = k, z_{t+1} = l\right) p\left(z_{t+1} = l \middle| z_t = k\right) \\ &= \textstyle\sum_{l=1}^{K} p\left(\mathbf{x}_{t+1:T} \middle| z_{t+1} = l\right) P_{k,l} \end{aligned}$$

and

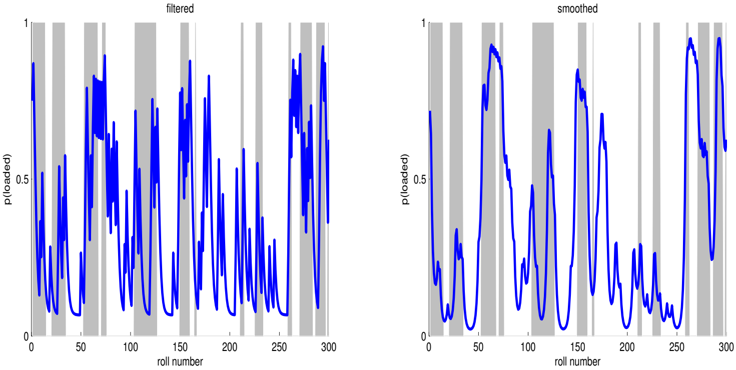
$$\begin{aligned} p\left(\mathbf{x}_{t:T} \middle| z_t = k\right) &= p\left(\mathbf{x}_t, \mathbf{x}_{t+1:T} \middle| z_t = k\right) \\ &= p_k\left(\mathbf{x}_t\right) p\left(\mathbf{x}_{t+1:T} \middle| z_t = k\right) \end{aligned}$$

Example: Casino

- In this model, $x_t \in \{1, 2, ..., 6\}$ represents which dice face shows up, and z_t represents the identity of the dice that is being used. Most of the time the casino uses a fair dice, $z = 1$, but occasionally it switches to a loaded dice, $z = 2$, for a short period.
- If $z = 1$ the observation distribution is a uniform distribution over $\{1, 2, ..., 6\}$. If $z = 2$, the observation distribution is skewed towards face 6.



Example: Casino



Inference in the dishonest casino. Vertical gray bars denote the samples that we generated using a loaded die. (left) Filtered estimate of probability of using a loaded dice. We hope to see a spike up whenever there is a gray bar. (right) Smoothed estimates.