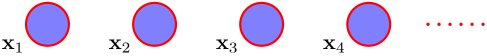
C.M. Bishop: Pattern Recognition and Machine Learning Ch. 13. Sequential data

Mari-Sanna Paukkeri

April 23, 2007

Outline

Introduction



- Sets of data points assumed to be independent and identically distributed (i.i.d) so far
- i.i.d is a poor assumption for *sequential data*
  - measurements of time series (rainfall), daily values of a currency exchange rate, acoustic features in speech recognition
  - sequence of nucleotide base pairs along a strand of DNA, sequence of characters in an English sentence

Markov model

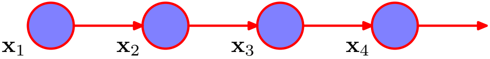
- Markov model:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) \tag{13.1}$$

- Each of the conditional distributions is independent of all previous observations except N most recent

The first-order Markov chain

- Homogeneous Markov chain



- Joint distribution for a sequence of N observations

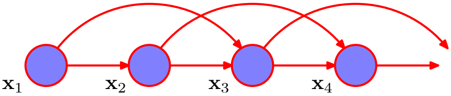
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n | \mathbf{x}_{n-1}) \qquad (13.2)$$

- From the d-separation property

$$p(\mathbf{x}_n | \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n | \mathbf{x}_{n-1}) \qquad (13.3)$$

A higher-order Markov chain

The second-order Markov chain



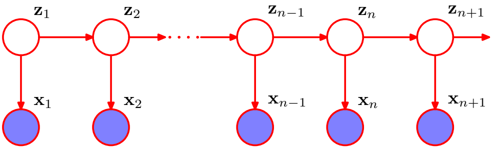
- The joint distribution

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1)\prod_{n=3}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}) \qquad (13.4)$$

A higher-order Markov chain

- Observations are discrete variables having K states
- first-order: $K - 1$ parameters for each K states $\rightarrow K(K-1)$ parameters
- Mth order: $K^{M-1}(K-1)$ parameters

Hidden Markov models (HMM)



- $\mathbf{z}_n$ latent variables (discrete)
- $\mathbf{x}_n$ observed variables
- The joint distribution of the state space model

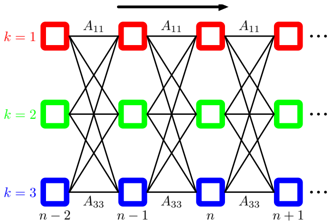
$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n | \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{z}_n) \tag{13.6}$$

Hidden Markov models (HMM)

- Transition probability

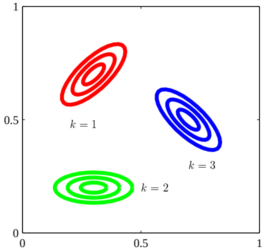
$$p(\mathbf{z}_n|\mathbf{z}_{n-1,\mathbf{A}}) = \prod_{k=1}^{K}\prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}}$$

$A_{jk} \equiv p(z_{nk} = 1|z_{n-1,j} = 1),$
$0 \leq A_{jk} \leq 1$ and $\sum_k A_{jk} = 1$



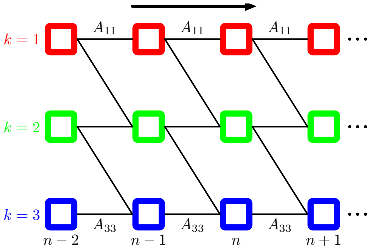
- Emission probability

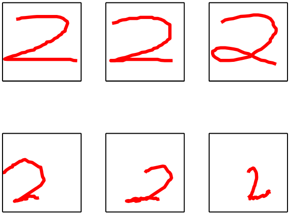
$$p(\mathbf{x}_n|\mathbf{z}_n, \phi) = \prod_{k=1}^{K} p(\mathbf{x}_n|\phi_k)^{z_{nk}}$$



HMM applications

- Speech recognition
- Natural language modelling
- Analysis of biological sequences (e.g. proteins and DNA)
- On-line handwriting recognition; Example: Handwritten digits
  - Left-to-right architecture
  - On-line data: each digit represented by the trajectory of the pen as a function of time





Maximum likelihood for the HMM

- We have observed a data set

$$\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\},$$

- so we can determine the parameters of an HMM

$$\theta = \{\pi, \mathbf{A}, \phi\}$$

  by using maximum likelihood.

- The likelihood function is

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \tag{13.11}$$

Maximizing the likelihood function

Expectation maximization algorithm (EM)

- Initial selection for the model parameters: θ^{old}
- E step:
  - Posterior distribution of the latent variables $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$

$$Q(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) \qquad (13.12)$$

Maximizing the likelihood function: EM

E step:

$$Q(\theta, \theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(\mathbf{x}_n|\phi_k) \quad (13.17)$$

- The marginal posterior distribution of a latent variable γ and the joint posterior distribution of two successive latent variables ξ

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}, \theta^{\text{old}}) \quad (13.13)$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_{n-1}, \mathbf{z}_n|\mathbf{X}, \theta^{\text{old}}) \quad (13.14)$$

Maximizing the likelihood function: EM

M step:

- Maximize $Q(\theta, \theta^{\text{old}})$ with respect to parameters $\theta = \{\pi, \mathbf{A}, \phi\}$, treat $\gamma(\mathbf{z}_n)$ and $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$ as constant. By using Lagrange multipliers

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \tag{13.18}$$

$$A_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1,j}, z_{nl})} \tag{13.19}$$

Maximizing the likelihood function: EM

M step:

- Parameters ϕ_k independent
  $\longrightarrow$ for Gaussian emission densities $p(\mathbf{x}|\phi_k) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad (13.20)$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nk})} \qquad (13.21)$$

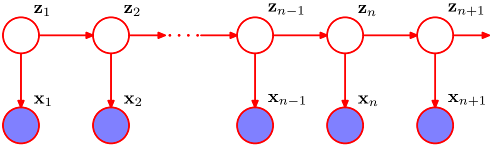
Back to the problem...

- We have observed a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,
- so we can determine the parameters of an HMM $\theta = \{\pi, \mathbf{A}, \phi\}$
- by maximizing the likelihood function $p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$.

- We used EM to maximize $Q(\theta, \theta^{\text{old}})$ and resulted to coefficients $\pi_k(\gamma)$, $A_{jk}(\xi)$, $\mu_k(\gamma)$ and $\Sigma_k(\gamma)$.

- How to evaluate γ and ξ?

The forward-backward algorithm



- Two-stage message passing algorithm
- Several variants, we focus on alpha-beta algorithm

Evaluate $\gamma(\mathbf{z}_n)$

- Using Bayes' theorem

$$\gamma(\mathbf{z}_n) = p(\mathbf{z}_n|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{z}_n)p(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.32}$$

$$= \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n)p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n)}{p(\mathbf{X})}$$

$$= \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.33}$$

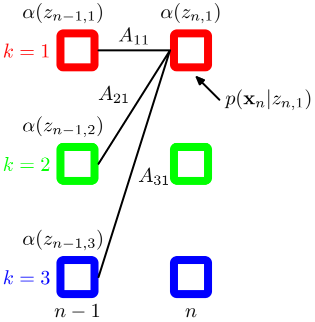
- where we have defined

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_n) \tag{13.34}$$

$$\beta(\mathbf{z}_n) = p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{z}_n) \tag{13.35}$$

Evaluate $\gamma(\mathbf{z}_n)$: forward-backward

Forward recursion for $\alpha(\mathbf{z}_n)$

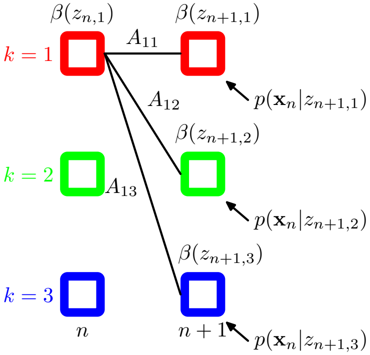


$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1}) \tag{13.36}$$

$$\alpha(\mathbf{z}_1) = p(\mathbf{x}_1, \mathbf{z}_1) = p(\mathbf{z}_1) p(\mathbf{x}_1|\mathbf{z}_1) = \prod_{k=1}^{K} \{\pi_k p(\mathbf{x}_1|\phi_k)\}^{z_{1k}} \tag{13.37}$$

Evaluate $\gamma(\mathbf{z}_n)$: forward-backward

Backward recursion for $\beta(\mathbf{z}_n)$



$$\beta(\mathbf{z}_n) = \sum_{\mathbf{z}_{n+1}} \beta(\mathbf{z}_{n+1}) p(\mathbf{x}_{n+1}|\mathbf{z}_{n+1}) p(\mathbf{z}_{n+1}|\mathbf{z}_n) \tag{13.38}$$

$$\beta(\mathbf{z}_N) = 1$$

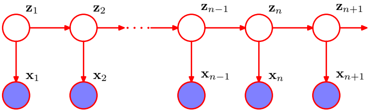
Evaluate $\xi(\mathbf{z}_{n-1}, \mathbf{z}_n)$

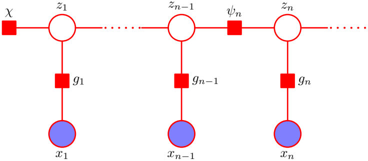
- Using Bayes' theorem

$$\begin{aligned}
\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) &= p(\mathbf{z}_{n-1}, \mathbf{z}_n | \mathbf{X}) \\
&= \frac{p(\mathbf{X}|\mathbf{z}_{n-1}, \mathbf{z}_n) p(\mathbf{z}_{n-1}, \mathbf{z}_n)}{p(\mathbf{X})} \\
&= \frac{\alpha(\mathbf{z}_{n-1}) p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{z}_n | \mathbf{z}_{n-1}) \beta(\mathbf{z}_n)}{p(\mathbf{X})} \qquad (13.43)
\end{aligned}$$

The sum-product algorithm for the HMM

- Solve the problem of finding local marginals for the hidden variables γ and ξ
- Can be used instead of forward-backward algorithm





- Results in

$$\gamma(\mathbf{z}_n) = \frac{\alpha(\mathbf{z}_n)\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.54}$$

$$\xi(\mathbf{z}_{n-1}, \mathbf{z}_n) = \frac{\alpha(\mathbf{z}_{n-1})p(\mathbf{x}_n|\mathbf{z}_n)p(\mathbf{z}_n|\mathbf{z}_{n-1})\beta(\mathbf{z}_n)}{p(\mathbf{X})} \tag{13.43}$$

Scaling factors

- Used to solve forward-backward algorithm

$$\alpha(\mathbf{z}_n) = p(\mathbf{x}_n|\mathbf{z}_n) \sum_{\mathbf{z}_{n-1}} \alpha(\mathbf{z}_{n-1}) p(\mathbf{z}_n|\mathbf{z}_{n-1}) \tag{13.36}$$

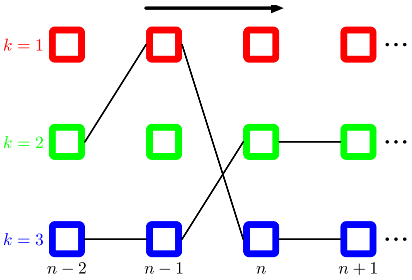
- Probabilities $p(\mathbf{x}_n|\mathbf{z}_n)$ and $p(\mathbf{z}_n|\mathbf{z}_{n-1})$ are often significantly less than unity
  $\rightarrow$ values $\alpha(\mathbf{n}_n)$ go to zero exponentially quickly
- We introduce re-scaled versions

$$\hat{\alpha}(\mathbf{z}_n) = \frac{\alpha(\mathbf{z}_n)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_n)} \tag{13.55}$$

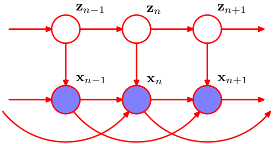
$$\hat{\beta}(\mathbf{z}_n) = \frac{\beta(\mathbf{z}_n)}{p(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N|\mathbf{x}_1, \ldots, \mathbf{x}_n)}$$

The Viterbi algorithm

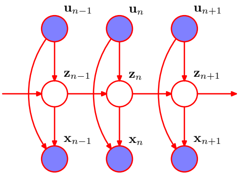
- Finding the most probable sequence of latent states is not the same as that of finding the set of states that are individually the most probable.
  - The latter problem has been solved already
  - The max-sum algorithm (Viterbi algorithm) can be used to solve the former problem



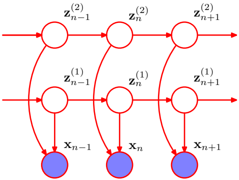
Extensions of the hidden Markov model



Autoregressive HMM

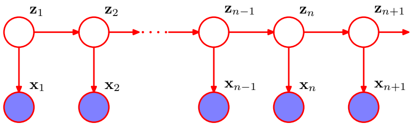


Input-output HMM



Factorial HMM

Linear Dynamical Systems



A linear-Gaussian model

- The general form of algorithms for the LDS are the same as for the HMM
- Continuous latent variables
- Both observed $\mathbf{x}_n$ and latent $\mathbf{z}_n$ variables Gaussian
  - Joint distribution over all variables, marginals and conditionals are Gaussian

  $\Rightarrow$ The sequence of individually most probable latent variable values is the same as the most probable latent sequence (no Viterbi considerations)

Linear Dynamical Systems

- Transition and emission probabilities

$$p(\mathbf{z}_n|\mathbf{z}_{n-1}) = \mathcal{N}(\mathbf{z}_n|\mathbf{A}\mathbf{z}_{n-1}, \Gamma) \tag{13.75}$$
$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\mathbf{x}_n|\mathbf{C}\mathbf{z}_n, \Sigma) \tag{13.76}$$

- The initial latent variable

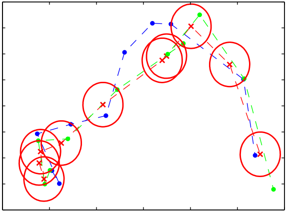
$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{z}_1|\mu_0, \mathbf{V}_0) \tag{13.77}$$

- The parameters $\theta = \{\mathbf{A}, \Gamma, \mathbf{C}, \Sigma, \mu_0, \mathbf{V}_0\}$ determined using maximum likelihood through EM

Inference in LDS

1. Find the marginal distributions for the latent variables conditional on the observation sequence
2. Given the parameters $\theta = \{\mathbf{A}, \Gamma, \mathbf{C}, \Sigma, \mu_0, \mathbf{V}_0\}$, predict the next latent state $\mathbf{z}_{n+1}$ and next observation $\mathbf{x}_{n+1}$

- Sum-product algorithm
  - Kalman filter (forward-recursion, α message)
  - Kalman smoother (backward-recursion, β message)
- Application of the Kalman filter: tracking



- True positions of the object
- Noisy measurements of the positions
- × Means of the inferred positions

Learning in LDS

- Determine $\theta = \{\mathbf{A}, \Gamma, \mathbf{C}, \Sigma, \mu_0, \mathbf{V}_0\}$ using maximum likelihood (again)
- Expectation maximization
    - E step:

$$Q(\theta, \theta^{\text{old}}) = \mathbb{E}_{\mathbf{Z}|\theta^{\text{old}}}[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] \tag{13.109}$$

    - M step: Maximize with respect to the components of θ

Extensions of LDS

- The marginal distribution of the observed variables is Gaussian
  $\Rightarrow$ use *Gaussian mixture* as the initial distribution for $\mathbf{z}_1$
- Make Gaussian approximation by linearizing around the mean of the predicted distribution
  - *Extended Kalman filter*
- Combining the HMM with a set of linear dynamical systems
  - *Switching state space model*

Particle filters

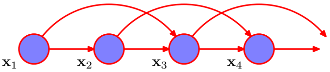
Sampling methods

- Needed for dynamical systems which do not have a linear-Gaussian
- Sampling-importance-resampling formalism
  $\Rightarrow$ a sequential Monte Carlo as the particle filter
- Particle filter algorithm:
  At time step n
  - obtained a set of samples and weights
  - observe $\mathbf{x}_{n+1}$
  - evaluate samples and weights for time step $n+1$

Summary

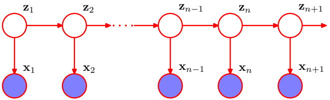
Markov model

- Discrete observed variables; each depends on N previous observations



Hidden Markov model

- Discrete latent variables



Linear dynamical systems

- Continuous latent variables