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

Mari-Sanna Paukkeri

April 23, 2007
Introduction

Markov Models

Hidden Markov Models
- Maximum likelihood for the HMM
- The forward-backward algorithm
- The sum-product algorithm for the HMM
- Scaling factors
- The Viterbi algorithm
- Extensions of the hidden Markov model

Linear Dynamical Systems
- Inference in LDS
- Learning in LDS
- Extensions of LDS
- Particle filters

Summary
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:

\[
p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n|x_1, \ldots, x_{n-1}) \quad (13.1)
\]

- 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

\[ p(x_1, \ldots, x_N) = p(x_1) \prod_{n=2}^{N} p(x_n|x_{n-1}) \]  

(13.2)

- From the d-separation property

\[ p(x_n|x_1, \ldots, x_{n-1}) = p(x_n|x_{n-1}) \]  

(13.3)
A higher-order Markov chain

The second-order Markov chain

The joint distribution

\[ p(x_1, \ldots, x_N) = p(x_1)p(x_2|x_1) \prod_{n=3}^{N} p(x_n|x_{n-1}, x_{n-2}) \]  \hspace{1cm} (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
- \( M \)th order: \( K^{M-1}(K - 1) \) parameters
Hidden Markov models (HMM)

- $z_n$ latent variables (discrete)
- $x_n$ observed variables

The joint distribution of the state space model

$$p(x_1, \ldots, x_N, z_1, \ldots, z_N) = p(z_1) \left[ \prod_{n=2}^{N} p(z_n | z_{n-1}) \right] \prod_{n=1}^{N} p(x_n | z_n)$$

(13.6)
Hidden Markov models (HMM)

- Transition probability

\[ p(z_n|z_{n-1}, A) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j}z_{nk}} \]

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

- Emission probability

\[ p(x_n|z_n, \phi) = \prod_{k=1}^{K} p(x_n|\phi_k)^{z_{nk}} \]
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
We have observed a data set

\[ X = \{x_1, \ldots, x_N\}, \]

so we can determine the parameters of an HMM

\[ \theta = \{\pi, A, \phi\} \]

by using maximum likelihood.

The likelihood function is

\[ p(X|\theta) = \sum_Z p(X, Z|\theta) \]  \hspace{1cm} (13.11)
Expectation maximization algorithm (EM)

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

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta)$$  \hspace{1cm} (13.12)
Maximizing the likelihood function: EM

E step:

\[ Q(\theta, \theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi(z_{n-1,j}, z_{nk}) \ln A_{jk} \]

\[ + \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln p(x_n|\phi_k) \]  

(13.17)

- The marginal posterior distribution of a latent variable \( \gamma \) and the joint posterior distribution of two successive latent variables \( \xi \)

\[ \gamma(z_n) = p(z_n|X, \theta^{\text{old}}) \]  

(13.13)

\[ \xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X, \theta^{\text{old}}) \]  

(13.14)
Maximizing the likelihood function: EM

M step:

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

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

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

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

\[
\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{13.20}
\]

\[
\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{13.21}
\]
Back to the problem...

- We have observed a data set $X = \{x_1, \ldots, x_N\}$,
- so we can determine the parameters of an HMM $\theta = \{\pi, A, \phi\}$
- by maximizing the likelihood function $p(X|\theta) = \sum_Z p(X, Z|\theta)$.

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

- How to evaluate $\gamma$ and $\xi$?
The forward-backward algorithm

- Two-stage message passing algorithm
- Several variants, we focus on alpha-beta algorithm
Hidden Markov Models
The forward-backward algorithm

Evaluate $\gamma(z_n)$

- Using Bayes’ theorem

$$\gamma(z_n) = p(z_n | X) = \frac{p(X | z_n) p(z_n)}{p(X)}$$  \hspace{1cm} (13.32)

$$= \frac{p(x_1, \ldots, x_n, z_n)p(x_{n+1}, \ldots, x_N | z_n)}{p(X)}$$

$$= \frac{\alpha(z_n) \beta(z_n)}{p(X)}$$  \hspace{1cm} (13.33)

- where we have defined

$$\alpha(z_n) = p(x_1, \ldots, x_n, z_n)$$  \hspace{1cm} (13.34)

$$\beta(z_n) = p(x_{n+1}, \ldots, x_N | z_n)$$  \hspace{1cm} (13.35)
Hidden Markov Models

The forward-backward algorithm

Evaluate $\gamma(z_n)$: forward-backward

Forward recursion for $\alpha(z_n)$

$$\alpha(z_n) = p(x_n|z_n) \sum_{z_{n-1}} \alpha(z_{n-1}) p(z_n|z_{n-1})$$ \hspace{1cm} (13.36)

$$\alpha(z_1) = p(x_1, z_1) = p(z_1) p(x_1|z_1) = \prod_{k=1}^{K} \left\{ \pi_k p(x_1|\phi_k) \right\}^{z_{1k}}$$ \hspace{1cm} (13.37)
The forward-backward algorithm

**Evaluate** $\gamma(z_n)$: forward-backward

Backward recursion for $\beta(z_n)$

$$\beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1}|z_{n+1}) p(z_{n+1}|z_n) \quad (13.38)$$

$$\beta(z_N) = 1$$
Using Bayes’ theorem

\[
\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X) = \frac{p(X|z_{n-1}, z_n)p(z_{n-1}, z_n)}{p(X)} = \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(X)}
\] (13.43)
The sum-product algorithm for the HMM

- Solve the problem of finding local marginals for the hidden variables $\gamma$ and $\xi$
- Can be used instead of forward-backward algorithm

Results in

$$
\gamma(z_n) = \frac{\alpha(z_n)\beta(z_n)}{p(X)} \quad (13.54)
$$

$$
\xi(z_{n-1}, z_n) = \frac{\alpha(z_{n-1})p(x_n|z_n)p(z_n|z_{n-1})\beta(z_n)}{p(X)} \quad (13.43)
$$
Scaling factors

- Used to solve forward-backward algorithm

\[ \alpha(z_n) = p(x_n|z_n) \sum_{z_{n-1}} \alpha(z_{n-1}) p(z_n|z_{n-1}) \]  

(13.36)

- Probabilities \( p(x_n|z_n) \) and \( p(z_n|z_{n-1}) \) are often significantly less than unity
  \[ \rightarrow \text{values } \alpha(n_n) \text{ go to zero exponentially quickly} \]

- We introduce re-scaled versions

\[ \hat{\alpha}(z_n) = \frac{\alpha(z_n)}{p(x_1, \ldots, x_n)} \]  

(13.55)

\[ \hat{\beta}(z_n) = \frac{\beta(z_n)}{p(x_{n+1}, \ldots, x_N|x_1, \ldots, x_n)} \]
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
A linear-Gaussian model

- The general form of algorithms for the LDS are the same as for the HMM
- Continuous latent variables
- Both observed $x_n$ and latent $z_n$ variables Gaussian
  - Joint distribution over all variables, marginals and conditionals are Gaussian
  - 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

\[ p(z_n|z_{n-1}) = \mathcal{N}(z_n|Az_{n-1}, \Gamma) \quad (13.75) \]
\[ p(x_n|z_n) = \mathcal{N}(x_n|Cz_n, \Sigma) \quad (13.76) \]

- The initial latent variable

\[ p(z_1) = \mathcal{N}(z_1|\mu_0, V_0) \quad (13.77) \]

- The parameters \( \theta = \{A, \Gamma, C, \Sigma, \mu_0, V_0\} \) 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 $\theta = \{A, \Gamma, C, \Sigma, \mu_0, V_0\}$, predict the next latent state $z_{n+1}$ and next observation $x_{n+1}$.

- Sum-product algorithm
  - Kalman filter (forward-recursion, $\alpha$ message)
  - Kalman smoother (backward-recursion, $\beta$ 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 $\theta = \{A, \Gamma, C, \Sigma, \mu_0, V_0\}$ using maximum likelihood (again)
- Expectation maximization
  - $E$ step:

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

  - $M$ step: Maximize with respect to the components of $\theta$
The marginal distribution of the observed variables is Gaussian
⇒ use Gaussian mixture as the initial distribution for $z_1$
- 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
  ⇒ a sequential Monte Carlo as the particle filter
- Particle filter algorithm:
  At time step $n$
  - obtained a set of samples and weights
  - observe $x_{n+1}$
  - 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