Introduction to Statistical Machine Learning

Christfried Webers

Statistical Machine Learning Group
NICTA
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2013

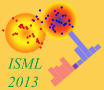
(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Outlines

Part XXI

Sequential Data 1

Sequential Data



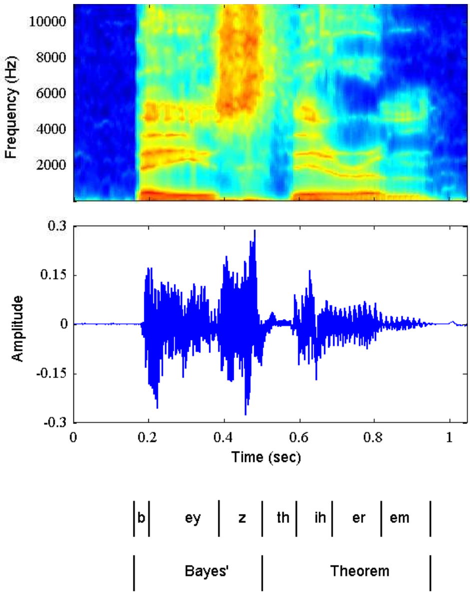
- For many applications, the i.i.d. assumption is a poor one
  - Time series (currencies, rainfall, speech)
  - Sequence of nucleotide base pairs along a DNA strand
  - Sequence of characters in a sentence.
  - ...

A Current data may not be independent of previous data.

B The distribution of the data may change while the data of the sequence are drawn/produced/emitted.

- Both A and B.
- Note: Use 'past' and 'future' to describe an order on the observations, but the concept of sequence is not restricted to temporal sequences.

Sequential Data - Example



Stationary versus Nonstationary Sequential Distributions

- Stationary case : Data evolves in time, but the distribution from which it is drawn stays the same.
- Nonstationary case : The generative distribution changes itself with time.

We will focus on the stationary case.

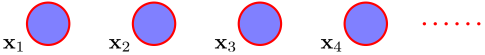
How the future depends on the past

- Goal: Predict the next value in a sequence.
- Assumption: Not all previous data are equally influencing the next value. (Technical problem: How to store an ever growing history of observations?)
- Assume that recent observations are more likely to be informative for the prediction of the next value than more historical observations.
- Is this always a good assumption?

Simplest Approach for Modelling a Sequence

©2013
Christfried Webers
NICTA
The Australian National University

- Treat all data as i.i.d.
- Example: Binary variable recording whether it rained or not on a day.
- Only information from such a model: frequency of rainfall.
- Can only use the frequency to predict whether it rains tomorrow or not. (Maybe not so bad for Canberra ;-)
- Usually: Observing whether it rained today helps to predict the weather for tomorrow.
- Need to relax the i.i.d. assumption to grasp this idea.



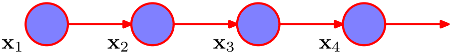
Markov Model for Modelling a Sequence

- One of the simplest ways to relax the i.i.d. assumption.
- Use the product rule to exactly express the joint distribution

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n \mid \mathbf{x}_1, \dots, \mathbf{x}_{n-1})$$

- Assume that each of the conditional expressions depends only on the most recent.
- First-order Markov chain

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n \mid \mathbf{x}_{n-1})$$



Markov Model for Modelling a Sequence

- Given the factorisation of the first-order Markov chain

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

- What is the conditional distribution for observation $\mathbf{x}_n$ given the previous observations?

©2013 Christfried Webers NICTA The Australian National University

Markov Model for Modelling a Sequence

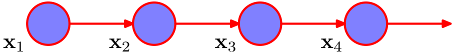
- Given the factorisation of the first-order Markov chain

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n \mid \mathbf{x}_{n-1})$$

- What is the conditional distribution for observation $\mathbf{x}_n$ given the previous observations?

$$\begin{aligned} p(\mathbf{x}_n \mid \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) &= \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_n)}{p(\mathbf{x}_1, \dots, \mathbf{x}_{n-1})} \\ &= \frac{p(\mathbf{x}_1) \prod_{i=2}^{n} p(\mathbf{x}_i \mid \mathbf{x}_{i-1})}{p(\mathbf{x}_1) \prod_{j=2}^{n-1} p(\mathbf{x}_j \mid \mathbf{x}_{j-1})} \\ &= p(\mathbf{x}_n \mid \mathbf{x}_{n-1}) \end{aligned}$$

(Not surprisingly, this matches our assumption.)

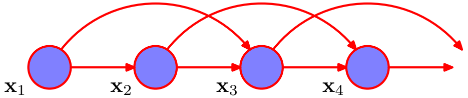


Second-order Markov Chain

©2013 Christfried Webers NICTA The Australian National University

- Assume that the trend in previous observations provides important information in predicting the next value.
- For a trend, we need at least two previous observations.

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



M^{th}-order Markov Chain

©2013 Christfried Webers NICTA The Australian National University

- Extend this idea to an M^{th}-order Markov chain in which the probability of each observation depends on the previous M observations

$$
\begin{aligned}
p(\mathbf{x}_1, \dots, \mathbf{x}_N) = & p(\mathbf{x}_1)\, p(\mathbf{x}_2 \mid \mathbf{x}_1) \dots p(\mathbf{x}_M \mid \mathbf{x}_1, \dots \mathbf{x}_{M-1}) \\
& \times \prod_{n=M+1}^{N} p(\mathbf{x}_n \mid \mathbf{x}_{n-1}, \mathbf{x}_{n-2} \dots \mathbf{x}_{n-M})
\end{aligned}
$$

- What is a reasonable M ?

M^{th}-order Markov Chain

- Assuming K different states for each variable $\mathbf{x}$, how many parameters does a M^{th}-order Markov chain have?
- $M = 0$: no Markov parameter, i.i.d. data
- $M = 1$: First-order Markov chain, $K - 1$ parameters for each of the K states of the previous observation. Number of parameters: $K(K - 1)$.
- M : M^{th}-order Markov Chain, $K - 1$ parameters for each of the K states of the previous M obervation. Number of parameters: $K^M(K - 1)$.
- Number of parameters grows exponentially with the order of the Markov chain. Impractical for larger M.

©2013 Christfried Webers NICTA The Australian National University

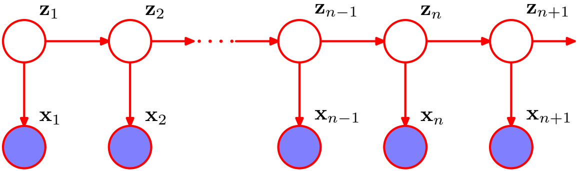
Extending the Model

- Goal: We want a model which is NOT restricted to the Markov assumption to any order. BUT can be specified by a limited number of free parameters.
- Use the idea of latent variables to construct a rich class of models out of simple components.
- (Remember mixture of Gaussians.)

State Space Model

- For each observation $\mathbf{x}_n$, a latent variable $\mathbf{z}_n$ is added.
- The type and dimensionality of $\mathbf{z}_n$ can differ from $\mathbf{x}_n$.
- Assume that the latent variables form a Markov chain.
- Key property: conditional independence of the latent variables

$$\mathbf{z}_{n+1} \perp\!\!\!\perp \mathbf{z}_{n-1} \mid \mathbf{z}_n$$

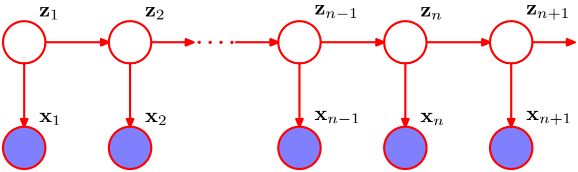


State Space Model

- Joint distribution for the state space model

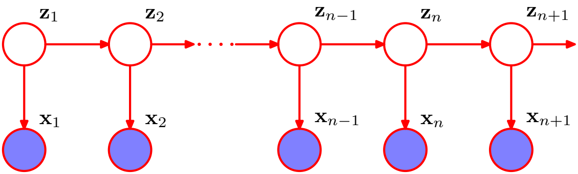
$$p(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{z}_1, \dots, \mathbf{z}_N) = p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n)$$

- d-separation of the graphical model: no blocked path between two observed variables $\mathbf{x}_i$ and $\mathbf{x}_j$ (no HT/TH node on any $\mathbf{x}_i$, nor any HH node on the latent variables $\mathbf{z}_i$.)
- Predictive distribution $p(\mathbf{x}_{n+1} \mid \mathbf{x}_1, \dots, \mathbf{x}_n)$ depends on all previous observations.
- The observed variables $\mathbf{x}_n$ do not satisfy the Markov property of any order.



State Space Model

- Two important models described by this graphical model.
- Hidden Markov Model or HMM : Latent variables $\mathbf{z}_n$ are discrete.
- Linear Dynamical System : Latent and observed variables are Gaussian with a linear-Gaussian dependence of the conditional distribution on their parents.



Hidden Markov Model

- State space model with discrete latent variables.

$$
\begin{aligned}
p(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{z}_1, \dots, \mathbf{z}_N) &= p(\mathbf{z}_1) \left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) \right] \prod_{n=1}^{N} p(\mathbf{x}_n \mid \mathbf{z}_n) \\
&= p(\mathbf{z}_1)\, p(\mathbf{x}_1 \mid \mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1})\, p(\mathbf{x}_n \mid \mathbf{z}_n)
\end{aligned}
$$

- Each step can be viewed as an extension of the mixture distribution model where each of the component densities is $p(\mathbf{x} \mid \mathbf{z})$. Choice of mixture component depends now on the previous state and is represented by $p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$.
- Latent variables are discrete multinomial variables $\mathbf{z}_n$ describing which component of the mixture is responsible for generating the corresponding observable $\mathbf{x}_n$.

Hidden Markov Model

- Assume 1-in-K coding scheme.
- Each latent variable $\mathbf{z}_n$ has K different states.
- The conditional distribution $p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$ is a table (matrix) with $K \times K$ entries, denoted by $\mathbf{A} \in \{0, 1\}^{K \times K}$.
- The elements of $\mathbf{A}$ are called transition probabilities

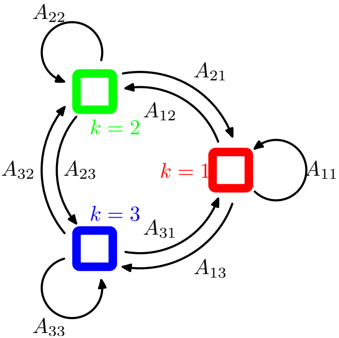
$$A_{jk} = p(z_{n,k} = 1 \mid z_{n-1,j} = 1).$$

  satisfying $0 \leq A_{jk} \leq 1$ and $\sum_{k=1}^{K} A_{jk} = 1$.
- The number of independent parameters is $K(K-1)$.

Hidden Markov Model - Transition Diagram

- Given the state $\mathbf{z}_{n-1}$ at step $n-1$, how what is the next state $\mathbf{z}_n$?

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

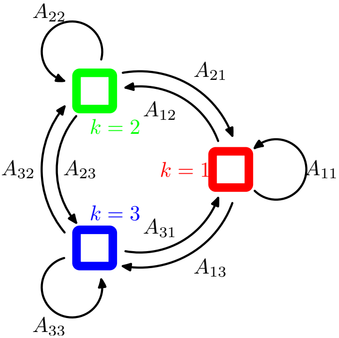


Transition Diagram for a model with three possible states.
Black lines denote the elements of the transition matrix A_{jk}.

Hidden Markov Model - Transition Diagram

- Initial latent node $\mathbf{z}_1$ has no parent, therefore the marginal distribution is given by a vector of probabilities $\boldsymbol{\pi}$ with $\pi_k = p(z_{1k} = 1)$ so that

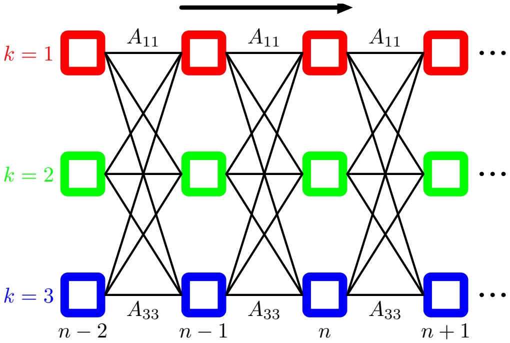
$$p(\mathbf{z}_1 \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \qquad \sum_k \pi_k = 1$$



Transition Diagram for a model with three possible states. Black lines denote the elements of the transition matrix A_{jk}.

HMM - Unfolded Transition Diagram

- Unfold the transition diagram over the steps to get a lattice, or trellis. (Note, the transitions $\mathbf{A}$ in each step can be different.)



Unfolded Transition Diagram for a model with 3 possible states. Each column corresponds to one latent variable $\mathbf{z}_n$.

Hidden Markov Model - Emission Probabilities

- Complete the HMM by defining the emission probabilities $p(\mathbf{x}_n \mid \mathbf{z}_n, \phi)$ where ϕ is a set of parameters governing the conditional distributions.
- As $\mathbf{x}_n$ is observed, and ϕ is given, $p(\mathbf{x}_n \mid \mathbf{z}_n, \phi)$ is a K-dimensional vector corresponding to the K possible states of the binary vector $\mathbf{z}_n$.
- Emission probabilities can be represented as

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

©2013
Christfried Webers
NICTA
The Australian National University

Hidden Markov Model - Homogeneous

- Remember: Transition probability in a Markov chain $T(\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{z}_n \mid \mathbf{z}_{n-1})$.
- A Hidden Markov Model is called homogeneous if the transition probabilities are the same for all steps n

$$p(\mathbf{z}_n \mid \mathbf{z}_{n-1}) = p(\mathbf{z}_{n-1} \mid \mathbf{z}_{n-2}) \qquad\qquad \forall\, n = 3, \dots, N$$

- Assume a homogeneous HMM in the following.

Homogeneous Hidden Markov Model

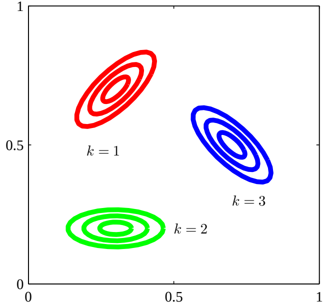
- Joint probability distribution over both latent and observed variables

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = p(\mathbf{z}_1 \mid \boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\mathbf{x}_m \mid \mathbf{z}_m, \phi)$$

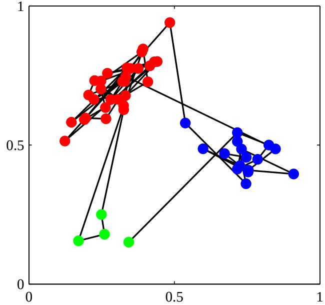
  where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)$, and $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \mathbf{A}, \phi\}$.
- Most of the discussion will be independent of the particular choice of emission probabilities (e.g. discrete tables, Gaussians, mixture of Gaussians).

Homogeneous HMM - Generative View

- Sampling from a hidden Markov model having a 3-state latent variable $\mathbf{z}$ and a Gaussian emission model $p(\mathbf{x} \mid \mathbf{z})$ where $\mathbf{x} \in \mathbb{R}^2$.



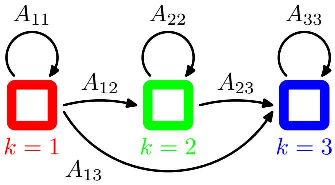
Contours of constant probability for the emission probabilities.



Sampling with 5% probability of change to each of the other states.

Hidden Markov Model - Variants

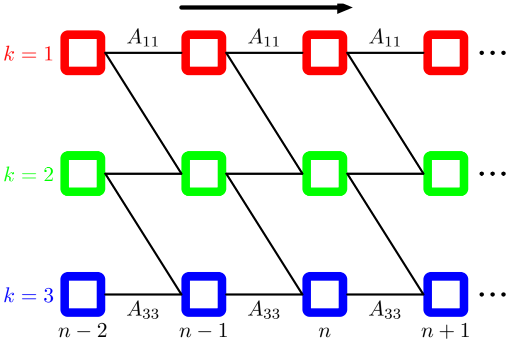
- Create variants of HMM by imposing constraints on the transition matrix $\mathbf{A}$.
- Left-to-right HMM : set all elements of $\mathbf{A}$ above the diagonal to zero, $A_{jk} = 0$ for $k < j$.
- Set the initial state probability $p(z_{11}) = 1$ and $p(z_{1j}) = 0$ for $j \neq 1$.
- Then, every sequence is constrained to start in state $k = 1$ and every state left can not be revisited again.



©2013
Christfried Webers
NICTA
The Australian National University

Hidden Markov Model - Variants

- Further restrict the transition matrix $\mathbf{A}$ to ensure that no large changes in the state space occur $A_{jk} = 0$ for $k > j + \Delta$ where Δ is the maximal change of state in one step.



Example of a HMM with $\Delta = 1$.

Hidden Markov Model - Handwritten Digits

- Digitize the trajectory online.
- $K = 16$ states representing a line segment (of fixed width) at 16 different angles.
- Trained with 45 examples of the digit ’2’.
- Left-to-right transmission probability with $\Delta = 1$.
- Upper row: Training samples used in 25 iterations of EM.
- Lower row: Sampled from the trained algorithm.