Algorithms for context prediction in Ubiquitous Systems

Markov prediction approaches

Stephan Sigg

Institute of Distributed and Ubiquitous Systems
Technische Universität Braunschweig

December 8, 2008

Overview and Structure

- Introduction to context aware computing
- Basics of probability theory
- Algorithms
  - Simple prediction approaches: ONISI and IPAM
  - Markov prediction approaches
  - The State predictor
  - Alignment prediction
  - Prediction with self organising maps
  - Stochastic prediction approaches: ARMA and Kalman filter
  - Alternative prediction approaohces
    - Dempster shafer
    - Evolutionary algorithms
    - Neural networks
    - Simulated annealing

Overview and Structure

- Introduction to context aware computing
- Basics of probability theory
- **Algorithms**
  - Simple prediction approaches: ONISI and IPAM
  - **Markov prediction approaches**
  - The State predictor
  - Alignment prediction
  - Prediction with self organising maps
  - Stochastic prediction approaches: ARMA and Kalman filter
  - Alternative prediction approahces
    - Dempster shafer
    - Evolutionary algorithms
    - Neural networks
    - Simulated annealing

Outline

Markov prediction approaches

Introduction and Markov properties

Historical remarks

- Markov processes
  - Intensively studied
  - Major branch in the theory of stochastic processes
- A. A. Markov (1856 – 1922)
- Extended by A. Kolmogorov by chains of infinitely many states
  - 'Anfangsgründe der Theorie der Markoffschen Ketten mit unendlich vielen möglichen Zuständen' (1936) [1]

[1] A. Kolmogorov, *Anfangsgründe der Theorie der Markoffschen Ketten mit unendlich vielen möglichen Zuständen*, 1936.

Introduction and Markov properties

Historical remarks

- Markov Chains
  - Theory of Markov chains applied to a variety of algorithmic problems
  - Standard tool in many probabilistic applications
- Intuitive graphical representation
  - Possible to illustrate properties of stochastic processes graphically
- Popular for their simplicity and easy applicability to huge set of problems[2]

[2] William Feller, *An introduction to probability theory and its applications*, Wiley, 1968.

Introduction and Markov properties

Introcution

- Dependent trials of events
  - Set of possible outcomes of a measurement E_i associated with occurence probability p_i
- When occurence of events is not independent
  - Probability to observe specific sequence $E_1, E_2, \dots, E_i$ obtained by conditional probability:

$$P(E_i|E_1, E_2, \dots, E_{i-1}) \tag{1}$$

  - In general:

$$P(E_i|E_1, E_2, \dots, E_{i-1}) \neq P(E_i|E_2, \dots, E_{i-1}) \tag{2}$$

Introduction and Markov properties

Independent random variables

- Sequence of tials for independent random variable
- T: number of trials up to first success of probability p.
- Then:

$$P\{T > k\} = (1 - p)^k \tag{3}$$

- Suppose: No success during the first m trials
- Waiting time T to first success for m-th trial has same distribution $(1 - p)^k$
- Independent of number of preceding failures m

Introduction and Markov properties

Examples

- Independent random variables
  - Number of coin tosses until 'head' is observed
  - Radioactive atoms always have the same probability of decaying at the next trial
- Dependent random variables
  - The knowledge that no streetcar has passed for five minutes increases our expectation that it will come soon.
  - Coin tossing:
    - Probability that the cumulative numbers of heads and tails will equalize at the second trial is $\frac{1}{2}$
    - Given that they did not, the probability that they equalize after two additional trials is only $\frac{1}{4}$

Introduction and Markov properties

Lack of memory – Rigorous formulation

- Suppose a waiting time T assumes the values $0, 1, 2, \dots$ with probabilities $p_0, p_1, p_2, \dots$
- Let T have the following property
  - Conditional probability that the waiting time terminates at the k-th trial equals p_0
- Then:
  - $p_k = (1 - p_0)^k p_0$

Introduction and Markov properties

Lack of Memory – Rigorous formulation

Proof.

- $1 - p_k = p_{k+1} + p_{k+2} + \cdots = P\{T > k\}$
- Conditional probability of $T = k$: $p_k/(1 - p_{k-1})$
- Assumption for all $k \geq 1$: $\frac{p_k}{1-p_{k-1}} = p_0$
- Since $p_k = (1 - p_{k-1}) - (1 - p_k)$

$$\frac{1 - p_k}{1 - p_{k-1}} = 1 - p_0 \tag{4}$$

- since $1 - p_0 = p_1 + p_2 + \ldots : 1 - p_k = (1 - p_0)^{k+1}$ and

$$p_k = 1 - p_{k-1} - (1 - p_k) = (1 - p_0)^k p_0 \tag{5}$$

□

Introduction and Markov properties

Markov property

Markov property

In the theory of stochasitc processes the described lack of memory is connected with the Markovian property.

Outline

Markov prediction approaches

Markov chains

Dependence and independence of events

- Independent trials of events
  - Set of possible outcomes of a measurement E_i associated with occurence probability p_i
- Probability to observe sample sequence:
  - $P\{(E_1, E_2, \ldots, E_i)\} = p_1 p_2 \cdots p_i$

Markov chains

Dependence and independence of events

- Theory of Markov chains:
  - Outcome of any trial depends exclusively on the outcome of the directly preceding trial
  - Outcome of E_k is no longer associated with fixed probability p_k
    - Instead: With every pair (E_i, E_j) a conditional probability p_{ij}
    - Probability that E_j is observed after E_i
    - Additionally: Probability a_i of the event E_i

Markov chains

Dependence and independence of events

- Theory of Markov chains:
  - $P\{(E_i, E_j)\} = a_i p_{ij}$
  - $P\{(E_i, E_j, E_k)\} = a_i p_{ij} p_{jk}$
  - $P\{(E_i, E_j, E_k, E_l)\} = a_i p_{ij} p_{jk} p_{kl}$
  - $P\{(E_i, E_j, \dots, E_m, E_n)\} = a_i p_{ij} \dots p_{mn}$

Markov chains

Markov chain

Markov chain

A sequence of observations $E_1, E_2, \ldots$ is called a Markov chain if the probabilities of sample sequences are defined by

$$P(E_1, E_2, \ldots, E_i) = a_1 \cdot p_{12} \cdot p_{23} \cdot \cdots \cdot p_{(i-1)i}. \tag{6}$$

and fixed conditional probabilities p_{ij} that the event E_i is observed directly in advance of E_j.

Markov chains

Markov chain

- Markov chain described by probability a for initial distribution and matrix P of transition probabilities.

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{7}$$

- P is a square matrix with non-negative entries that sum to 1 in each row.

Markov chains

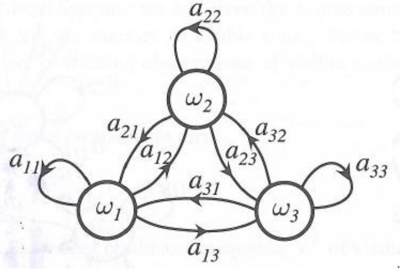
Stochastic matrix

- P is called a stochastic matrix.
- Any stochastic matrix is suited to describe transition probabilities of Markov chains.

Markov chains

Markov chain

- Markov chain sometimes modelled as directed graph $G = (V, E)$
- Labelled edges in E
- states (or vertices) in V.
- Transition probabilities p_{ij} between $E_i, E_j \in V$



Markov chains

Derive state transition probabilities

- p_{ij}^k denotes probability that E_j is observed exactly k observations after E_i was observed.
- Calculated as the sum of the probabilities for all possible paths $E_i E_{i_1} \cdots E_{i_{k-1}} E_j$ of length k
- We already know

$$p_{ij}^1 = p_{ij} \tag{8}$$

- Consequently:

$$P_{ij}^2 = \sum_{\nu} p_{i\nu} \cdot p_{\nu j} \tag{9}$$

Markov chains

Derive state transition probabilities

- By mathematical induction:

$$p_{ij}^{n+1} = \sum_{\nu} p_{i\nu} \cdot p_{\nu j}^{n} \tag{10}$$

- and

$$p_{ij}^{n+m} = \sum_{\nu} p_{i\nu}^{m} \cdot p_{\nu j}^{n} = \sum_{\nu} p_{i\nu}^{n} \cdot p_{\nu j}^{m} \tag{11}$$

Markov chains

Derive state transision probabilities

- Similar to the matrix P we can create a matrix P^n that contains all p_{ij}^n
- We obtain P_{ij}^{n+1} from P^{n+1} by multiplying all elements of the i-th row of P with the correspoinding elements of the j-the column of P^n and add all products.
- Symbolically: $P^{n+m} = P^n P^m$.

Markov chains
Examples

- Markov chains:
  - Urn models
    - Every Markov chain is equivalent to an urn model
    - Each urn represents a state in a markov chain and probabilities to draw specific balls represent possible events in this state
  - Branching processes
    - Instead of saying that the n-th trial results in E_k we say that the n-th generation is of size k
  - Random walk on a line
    - Events are transitions between states
    - Only two events are possible in each state

Markov chains

Random walks and ruin problems

- Random walk
  - When there are only two possible states E_1 and E_2 the matrix of transition probabilities is of the form

$$P = \begin{bmatrix} 1-p & p \\ \alpha & 1-\alpha \end{bmatrix} \tag{12}$$

- Can be realised by particle moving along one axis in one or the other direction.
- System is in state E_1 when the particle moves into one direction and in state E_2 otherwise.

Markov chains

Random walks and ruin problems

- Possible problems / questions
  - Expected time to return to origin
  - Expected time to return to origin given that the starting point had a specific distance to the origin
  - ...

Markov chains

Random walks and ruin problems

- Random walk with absorbing barriers

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 & 0 & 0 \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

- First and last state are obsorbing
- All inner states implement a random walk on the line
- Possible application: Game between two players with equal money balance where the loosing one has to pay one unit to the winner.

Markov chains

Random walks and ruin problems

- Random walk with reflecting barriers

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & \cdots & 0 & 0 & 0 \\ & & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1-p & p \end{bmatrix} \quad (14)$$

- First and last state are reflecting
- All inner states implement a random walk on the line

Markov chains

Random walks and ruin problems

Classical ruin problem

- Consider a gambler who wins or loses a dollar with probabilities p and $1 - p$
- Initial capital of gambler and adversary: z, $a - z$
- Game ends when the capital reaches 0 or a.
  - When one of the players is ruined
- We are interested in the probability of the gamblers ruin and the probability distribution of the game

Markov chains

Random walks and ruin problems

- Gamblers ruin problem
  - Random walk with absorbing barriers at 0 and a
- Examples:
  - Physicists use this model as crude approximation to one-dimensional diffusion or Brownian motion (Particle is exposed to great number of molecular collisions which impart to it a random motion)
  - $p > 1/2$ represents a drift to the right, when shocks from the left are more probable

Markov chains

Random walks and ruin problems

- Probability of gamblers ruin
  - q_z: Probability of gambler's ultimate ruin when z is the starting capital and a is the overal capital
  - After the first trial the gablers's fortune is either $z-1$ or $z+1$:
    - $q_z = pq_{z+1} + (1-p)q_{z-1}$
- We can show:

$$p \neq \frac{1}{2} \quad : \quad q_z = \frac{\left(\frac{1-p}{p}\right)^a - \left(\frac{1-p}{p}\right)^z}{\left(\frac{1-p}{p}\right)^a - 1} \tag{15}$$

$$p = \frac{1}{2} \quad : \quad q_z = 1 - \frac{z}{a} \tag{16}$$

Markov chains

Random walks and ruin problems

- Probability of gamblers ruin
  - The probability p_z of the gambler winning the game is equal to the probability of his adversary loosing the game.
  - It is therefore obtained in the same way by replacing p with $1 - p$ and z by $a - z$
  - Therefore: $p_z + q_z = 1$

Markov chains

Random walks and ruin problems

- Some interesting results
  - Since for $p = \frac{1}{2}$,we have derived $q_z = 1 - \frac{z}{a}$
    - A player with initial capital $z = 999$ has a probability of 0.999 to win a dollar before losing his capital.
  - With $p = 0.4$ the game is unfavorable, but still the probability of winning a dollar before losing the capital is about $\frac{2}{3}$

Markov chains

Random walks and ruin problems

Example – anecdote

A certain man used to visit Monte Carlo year after year and was always successful in recovering the cost of his vacations. He firmly believed in a magic power over chance.

This experience is not surprising.

- Assuming that he started with ten times the ultimate gain, the chances of success in any year are nearly 0.9.
- The probability of an unbroken sequence of ten successes is about $(1 - \frac{1}{10})^{10} \approx e^{-1} \approx 0.37$

Therefore, continued success is by no means improbable

- However: one failue would result in the gambler's ruin :-)

Markov chains

Random walks and ruin problems

ILLUSTRATING THE CLASSICAL RUIN PROBLEM

p	q	z	a	Probability of Ruin	Probability of Success	Expected Gain	Expected Duration
0.5	0.5	9	10	0.1	0.9	0	9
0.5	0.5	90	100	0.1	0.9	0	900
0.5	0.5	900	1,000	0.1	0.9	0	90,000
0.5	0.5	950	1,000	0.05	0.95	0	47,500
0.5	0.5	8,000	10,000	0.2	0.8	0	16,000,000
0.45	0.55	9	10	0.210	0.790	−1.1	11
0.45	0.55	90	100	0.866	0.134	−76.6	765.6
0.45	0.55	99	100	0.182	0.818	−17.2	171.8
0.4	0.6	90	100	0.983	0.017	−88.3	441.3
0.4	0.6	99	100	0.333	0.667	−32.3	161.7

The initial capital is z. The game terminates with ruin (loss z) or capital a (gain $a - z$).

- Effect of increasing stakes is more pronounced than might be expected

Markov chains

Random walks and ruin problems

- Expected duration of the game
  - D_z: Expected duration of the game when z is the starting capital and a is the overal capital
  - After the first trial the gablers's fortune is either $z-1$ or $z+1$:
    - $D_z = pD_{z+1} + (1-p)D_{z-1} + 1$
- We can show:

$$p \neq \frac{1}{2} \quad : \quad D_z = \frac{z}{1-2p} - \frac{a}{1-2p} \cdot \frac{1-\left(\frac{1-p}{p}\right)^z}{1-\left(\frac{1-p}{p}\right)^a} \tag{17}$$

$$p = \frac{1}{2} \quad : \quad D_z = z(a-z) \tag{18}$$

Markov chains

Random walks and ruin problems

- Expected duration of the game

$$p \neq \frac{1}{2} \quad : \quad D_z = \frac{z}{1-2p} - \frac{a}{1-2p} \cdot \frac{1-\left(\frac{1-p}{p}\right)^z}{1-\left(\frac{1-p}{p}\right)^a} \tag{19}$$

$$p = \frac{1}{2} \quad : \quad D_z = z(a-z) \tag{20}$$

Examples – Duration considerably longer as naively expected:

- If two players with 500 dollars each toss a fair coin, average duration of the game is 250000 trials
- If a gambler has only one dollar and his adversary 1000, the average duration is 1000 trials

Markov chains

Closures and closed sets

Closed set of states

- A set C of states is closed if no state outside C can be reached from any state E_i in C.
- For an arbitrary set C of states the smallest closest set containing C is called the closure of C
- A single state E_k forming a closed set is called absorbing

Markov chains

Closures and closed sets

Closed sets in stochastic matrices

If in a matrix P^n all rows and all columns corresponding to states outside a closed set C are deleted, the remaining matrices are again stochastic matrices.

Markov chains

Closures and closed sets

Irreducible Markov chain

A Markov chain is irreducible if there exists no closed set other than the set of all states.

Criterion for irreducible chains

A chain is irreducible if, and only if, every state can be reached from every other state.

Markov chains

Periodicity

Periodicity of states

- The state E_j has period $t > 1$ if $p_{jj}^n = 0$ unless $n = vt$ is a multiple of t and t is the largest integer with this property.
- the state E_j is aperiodic if no such $t > 1$ exists

- A state E_j to which no return i possible is considered aperiodic

Markov chains

Periodicity

- To deal with a periodic E_j it suffices to consider the chain at the trials $t, 2t, 3t$
- In this way we obtain a new Markov chain with transition probabilities p_{ik}^{t}
- In this new chain E_j is aperiodic
- Results concerning aperiodic states can thus be transferred to periodic states

Markov chains

Persistent and transient states

Persistent and transient states

- The state E_j is persistent if $\sum_{n=1}^{\infty} p_{jj}^{n} = 1$ and transient if $\sum_{n=1}^{\infty} p_{jj}^{n} < 1$
- A persistent state E_j is called null state if its mean recurrrence time $\mu_j = \infty$

Markov chains
Irreducible chains

- Two states are of the same type when they are either
  - both aperiodic
  - both have the same period
  - both are transient
  - both are persistent and each
    - with infinite recurrence times
    - or finite recurrence times

Markov chains

Irreducible chains

Type of states in irreducible chains

All states of an irreducible chain are of the same type

Markov chains

Applications – Card shuffling

- A deck of N cards can be arranged in $N!$ different orders.
- Each order represent a possible state of the system
- We conceive each particular shuffling operation as a transformation $E_i \rightarrow E_j$
- Result:
  - The permutation is not cyclic
  - Therefore, repeated application of a single operation will never visit all possible states
  - This means that the original state is again observed before all states are visited
- This is a Markov chain:
  - We assume that a player applies several shuffling operation with a random probability and that the current order of the cards is not known.

Markov chains

Markov processes

Markov process

A sequence of discrete-valued random variables is a markov process if the joint distribution of $(X^1, \ldots, X^n)$ is defined in such a way that the conditional probability of the relation $X^n = x$ on the hypothesis $X^{n_1} = x_1, \ldots, X^{n_r} = x_r$ is identical with the conditional probability of $X^n = x$ on the single hypothesis $X^{n_r} = x_r$.

Markov chains

Higher order Markov processes

- Order k Markov processes
- Typically
  - Occurence of event dependent on k events that were observed directly beforehand
  - Constrained lack of memory
  - Dependence between the last k events observed
- Useful for context prediction / time series forecasting, when typical patterns or trends are to be considered

Markov chains

Higher order Markov processes

- Probability that $E_1, E_2, \ldots, E_i$ observed is then

$$P(E_1, E_2, \ldots, E_i) = p_1 \cdot p_{12} \cdot p_{23} \cdot \cdots \cdot p_{(i-1)i}. \tag{21}$$

- Required: $p_i > 0 \forall i$ and $\sum_{p_i} = 1$.

Outline

Markov prediction approaches

Hidden Markov Models

Introduction

- Make a sequence of decisions for a process that is not directly observable[3]
- Current states of the process might be impacted by prior states
- HMM often utilised in speach recognition or gesture recognition

[3] Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern classification*, Wiley interscience, 2001.

Hidden Markov Models
Applications

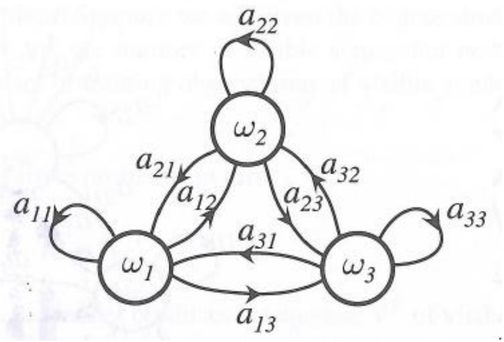
- Computational biology
  - Align biological sequences
  - Find sequences homologous to a known evolutionary family
  - Analyse RNA secondary structure [4]
- Computational linguistics[5]
  - Topic segmentation of text
  - Information extraction

[4] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*, Cambridge University Press, 1998.

[5] C.D. Manning and H. Schütze, *Foundations of statistical natural language processing*, MIT Press, 1999.

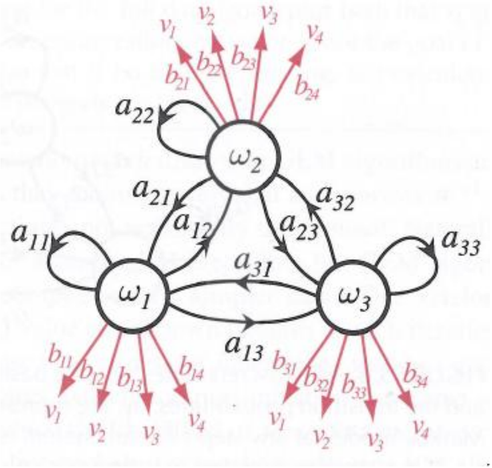
Hidden Markov Models

First order Markov models



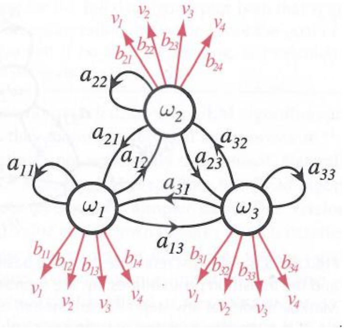
Hidden Markov Models

First order hidden Markov models



Hidden Markov Models

First order hidden Markov models



- At every time step t the system is in an internal state $\omega(t)$
- Additionally, we assume that it emits a (visible) symbol $v(t)$
- Only access to visible symbols and not to intenal states

Hidden Markov Models

First order hidden Markov models

- $V^T = \{v(1), v(2), \ldots, v(T)\}$
- In any state $\omega(t)$ we have a probability of emitting a particular visible symbol $v_k(t)$
- Probability to be in state $\omega_j(t)$ and emit symbol $v_k(t)$:
  - $P(v_k(t)|\omega_j(t)) = b_{jk}$
- Transmission probabilities: $p_{ij} = P(\omega_j(t+1)|\omega_i(t))$
- Emission probability: $b_{jk} = P(v_k(t)|\omega_j(t))$

Hidden Markov Models

First order hidden Markov models

- Normalisation conditions

$$\sum_j p_{ij} = 1 \text{for all } i \qquad (22)$$

$$\sum_k b_{jk} = 1 \text{for all } j \qquad (23)$$

Hidden Markov Models

First order hidden Markov models

- Central issues in hidden Markov models:
  - Evaluation problem
    - Determine the probability that a particular sequence of visible symbols V^T was generated by a given hidden Markov model
  - Decoding problem
    - Determine the most likely sequence of hidden states ω^T that led to a specific sequence of observations V^T
  - Learning problem
    - Given a set of training observations of visible symbols, determine the parameters p_{ij} and b_{jk} for a given HMM

Hidden Markov Models

First order hidden Markov models – Evaluation problem

- Probability that model produces a sequence V^T:

$$P(V^T) = \sum_{\overline{\omega}^T} P(V^T|\overline{\omega}^T)P(\overline{\omega}^T) \tag{24}$$

- Also:

$$P(\overline{\omega}^T) = \prod_{t=1}^{T} P(j\omega(t)|\omega(t-1)) \tag{25}$$

$$P(V^T|\overline{\omega}^T) = \prod_{t=1}^{T} P(v(t)|\omega(t)) \tag{26}$$

- Together:

$$P(V^T) = \sum_{\overline{\omega}^T} \prod_{t=1}^{T} P(v(t)|\omega(t))P(\omega(t)|\omega(t-1)) \tag{27}$$

Hidden Markov Models

First order hidden Markov models – Evaluation problem

- Probability that model produces a sequence V^T:

$$P(V^T) = \sum_{\overline{\omega}^T} \prod_{t=1}^{T} P(v(t)|\omega(t))P(\omega(t)|\omega(t-1)) \qquad (28)$$

- Formally complex but straightforward
- Naive computational complexity
  - $O(c^T T)$

Hidden Markov Models

First order hidden Markov models – Evaluation problem

- Probability that model produces a sequence V^T:

$$P(V^T) = \sum_{\overline{\omega}^T} \prod_{t=1}^{T} P(v(t)|\omega(t))P(\omega(t)|\omega(t-1)) \qquad (29)$$

- Computationally less complex algorithm:
  - Calculate $P(V^T)$ recursively
  - $P(v(t)|\omega(t))P(\omega(t)|\omega(t-1))$ involves only $v(t), \omega(t)$ and $\omega(t-1)$

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{ initial state} \\ 1 & t = 0 \text{ and } j = \text{ initial state} \\ \left[\sum_i \alpha_i(t-1)p_{ij}\right] b_{jk} v(t) & \text{otherwise} \end{cases} \qquad (30)$$

Hidden Markov Models

First order hidden Markov models – Evaluation problem

- Forward Algorithm
- Computational complexity: $O(c^2 T)$

Forward algorithm

```
initialise  t ← 0, p_ij, b_jk, V^T, α_j(o)
    for  t ← t + 1
        α_j(t) ← b_jk v(t) Σ_{i=1}^{c} α_i(t − 1) p_ij
    until  t = T
return  P(V^T) ← α_0(T)  for the final state
end
```

Hidden Markov Models

First order hidden Markov models – Decoding problem

- Given a sequence V^T, find the most probable sequence of hidden states.
- Enumeration of every possible path will cost $O(c^T T)$
  - Not feasible

Hidden Markov Models

First order hidden Markov models – Decoding problem

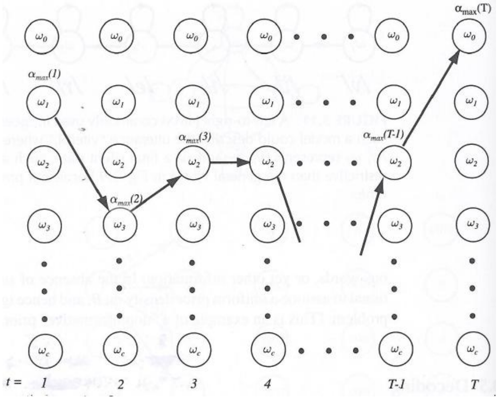
- Given a sequence V^T, find the most probable sequence of hidden states.

Decoding algorithm

initialise path $\leftarrow \{\}, t \leftarrow 0$
    for $t \leftarrow t+1$
        $j \leftarrow j+1$
        for $j \leftarrow j+1$
            $\alpha_j(t) \leftarrow b_{jk} v(t) \sum_{i=1}^{c} \alpha_i(t-1) p_{ij}$
        until $j = c$
        $j' \leftarrow arg\max_j \alpha_j(t)$
        append $\omega_{j'}$ to path
    until $t = T$
return path
end

Hidden Markov Models

First order hidden Markov models – Decoding problem



- Computational time of the decoding algorithm
  - $O(c^2 T)$
- However, computed path might be invalid

Hidden Markov Models

First order hidden Markov models – Learning problem

- Determine the model parameters p_{ij} and b_{jk}
  - Given: Training sample of observed values V^T
- No method known to obtain the optimal or most likely set of parameters from the data
  - However, we can nearly always determine a good solution by the forward-backward algorithm
  - General expectation maximisation algorithm
  - Iteratively update weights in order to better explain the observed training sequences

Hidden Markov Models

First order hidden Markov models – Learning problem

- Probability that the model is in state $\omega_i(t)$ and will generate the remainder of the given target sequence:

$$\beta_i(t) = \begin{cases} 0 & t = T \text{ and } \omega_i(t) \neq \omega_0 \\ 1 & t = T \text{ and } \omega_i(t) = \omega_0 \\ \sum_j \beta_j(t+1) p_{ij} b_{jk} v(t+1) & \text{otherwise} \end{cases} \tag{31}$$

Hidden Markov Models

First order hidden Markov models – Learning problem

- $\alpha_i(t)$ and $\beta_i(t)$ only estimates of their true values since transition probabilities p_{ij}, b_{jk} unknown
- Probability of transition between $\omega_i(t-1)$ and $\omega_j(t)$ can be estimated
    - Provided that the model generated the entire training sequence V^T by **any** path

$$\gamma_{ij}(t) = \frac{\alpha(t-1)p_{ij}b_{jk}\beta_j(t)}{P(V^T|\Theta)} \tag{32}$$

- Probability that model generated sequence V^T:

$$P(V^T|\Theta) \tag{33}$$

Hidden Markov Models

First order hidden Markov models – Learning problem

- Calculate improved estimate for p_{ij} and b_{jk}

$$\overline{p_{ij}} = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_{k} \gamma_{ik}(t)} \tag{34}$$

$$\overline{b_{jk}} = \frac{\sum_{t=1, v(t)=v_k}^{T} \sum_{l} \gamma_{jl}(t)}{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)} \tag{35}$$

- Start with rough estimates of p_{ij} and b_{jk}
- Calculate improved estimates
- Repeat until some conversion is reached

Hidden Markov Models

First order hidden Markov models – Lerning problem

Forward-Backward algorithm

initialise p_{ij}, b_{jk}, V^T, convergence criterion $\Theta, z \leftarrow 0$
    do $z \leftarrow z + 1$
        compute $\overline{p_{ij}(z)}$
        compute $\overline{b_{jk}(z)}$
        $p_{ij}(z) \leftarrow \overline{p_{ij}(z)}$
        $b_{jk}(z) \leftarrow \overline{b_{jk}(z)}$
    until $\max_{i,j,k}[p_{ij}(z) - p_{ij}(z-1), b_{jk}(z) - b_{jk}(z-1)] < \Theta$
        (convergence achieved)
return $p_{ij} \leftarrow p_{ij}(z)$, $b_{jk} \leftarrow b_{jk}(z)$
end

Hidden Markov Models

First order hidden Markov models

- Context prediction with hidden Markov models:
  1. Given the transition model, estimate all p_{ij} and b_{jk}
  2. Given a sequence V^T, decode the most probable sequence of hidden states
  3. Extrapolate the sequence of expected hidden states

Outline

Markov prediction approaches

1. Introduction and Markov properties
2. Markov chains
3. Hidden Markov Models
4. Context prediction with Markov approaches
   - Properties of Markov prediction approaches
5. Conditional random fields
   - Context prediction with CRF
   - Properties of CRF prediction approaches
6. Conclusion

Context prediction with Markov approaches

- Given: Sequence of contexts $\xi_{0-k+1}, \ldots, \xi_0$
- Generate Markov chain representing the transition probabilities for each pair of observations
- Now possible: Provide probability distribution on the next outcome
- Can also be generalised to higher order Markov processes
- Several iterations of this process provide higher prediction horizons

Properties of Markov prediction approaches

Memory and processing load

- Runtime dependent on size of probability graph G
- C: Set of different context values
- The number of states of the Markov chain: C.
- Time to find most probable next state is $O(|C|)$ in the worst case.
  - Every arc to possible following context to be considered.
  - $|C| - 1$ arcs existent in the worst case.
- Most probable n future context time series elements
  - Naive computation time: $O(|C|^n)$
  - When transition probabilities stored to a matrix, one matrix multiplication for every future context
  - Computation time: $O(n \cdot |C|^2)$.

Properties of Markov prediction approaches

Memory and processing load

- Memory requirements
  - Dependent on the number of contexts observed – size of the transition matrix
  - Order 1: $O(|C|^2)$
  - Order k: $O(|C|^{k+1})$

Properties of Markov prediction approaches

Prediction horizon

- Prediction horizon can be extended by iterative prediction
  - Utilise predicted contexts as input
- Problem: Less accurate
  - Predicted contexts more error prone than measured values

Properties of Markov prediction approaches
Adaptability

- The Markov prediction approach is well able to adapt to changing environments
    - Adapt context transition probabilities
    - Consideration of new events possible
        - Rebuild of transition matrix required

Properties of Markov prediction approaches

Multi-dimensional time series

- The Markov prediction algorithm is not suited for multi-dimensional time series
  - Designed for one-dimensional Input
  - Possible: Aggregation of multi-dimensional time series to one-dimensional time series.

Properties of Markov prediction approaches

Iterative prediction

- Iterative Prediction possible
  - Steep decrease in prediction accuracy expected since prediction horizon is only 1
  - Increase of prediction horizon possible by Aggregation of context sequence of fixed length in one Markov state
    - Prediction horizon fixed
    - Increase in Memory consumption and processing time
    - When l contexts are aggregated: l^C states
    - Runtime:
      $$O(n \cdot l^{C^2}).$$
    - Memory consumption:
      $$O(l^{C^2}) \text{ (order one)}$$
      $$O(l^{C^{k+1}}) \text{ (order k)}$$

Properties of Markov prediction approaches

Prediction of context durations

- Prediction of context duration not possible
  - Only simple sequence of occurring contexts possible

Properties of Markov prediction approaches

Approximate matching of patterns

- Exact pattern matching
  - The Markov prediction algorithm utilises exact pattern matching

Properties of Markov prediction approaches

Context data types

- All context data types supported
  - Every distinct context type one state
  - Probably drastic increase in runtime and memory consumption for numeric context types
  - Possible: Assign intervals to states

Properties of Markov prediction approaches

Pre-processing

- Pre-processing required to construct context transition probabilities
- On-line approach feasible – learning
- Runtime: $O(k)$
  - Count frequency of specific context transitions in training time series of length k

Aspects of prediction algorithms

Summary

	IPAM	ONISI	Markov	CRF
Numeric Contexts	yes	no	yes	
Non-numeric Contexts	yes	yes	yes	
Complexity	$O(k)$	( )	$O(C^2)$	
Learning ability	(no)	yes	yes	
Approximate matching	no	no	no	
Multi-dim. TS	(no)	(no)	(no)	
Discrete data	yes	yes	yes	
Variable length patterns	no	yes	no	
Multi-type TS	yes	no	(no)	
Continuous data	no	no	no	
Pre-processing	$O(k)$	–	$O(k)$	
Context durations	no	no	no	
Continuous time	no	no	yes	

Properties of Markov prediction approaches

Conclusion

- Markov processes are straightforward and easily applied to context prediction tasks.
- Model can be applied to numerical and non-numerical data alike.
- Prediction that reaches farther into future implicitly utilises already predicted data which might consequently decrease the prediction accuracy.

Outline

Markov prediction approaches

1. Introduction and Markov properties
2. Markov chains
3. Hidden Markov Models
4. Context prediction with Markov approaches
   - Properties of Markov prediction approaches
5. Conditional random fields
   - Context prediction with CRF
   - Properties of CRF prediction approaches
6. Conclusion

Conditional random fields

Introduction

- Undirected graphical model [6] [7]
- Similar to HMM
  - HMM specific CRF
  - Relax assumptions about input and output sequence
  - Instead of constant transition probability: Arbitrary functions that vary across positions in sequence of hidden states
- Vertices represent random variables
- Edges represent dependency between two random variables

[6] John Lafferty, Andrew McCallum and Fernando Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, In Proceedings of the 18th international conference on machine learning, pp 282-289, 2001.

[7] Douglas L. Vail, Manuela M. Veloso and John D. Lafferty, *Conditional random fields for activity recognition*, In Proceedings of the AAMAS, 2007.

Conditional random fields

Introduction

- Layout of inner-graph (hidden states) arbitrary
- Input sequence: X
- Inner states: Y
- Conditional dependency of each Y_i on X defined through set of feature functions

$$f(i, Y_{i-1}, Y_i, X) \tag{36}$$

  - Each feature assigned numerical weight

Conditional random fields

Training

- Various learning approaches to train CRF
  - Gradient based
  - Quasi-Newton-approach
- Sequences provided of which also desired output is known
- CRF parameters are adapted to match a maximum number of training sequences

Conditional random fields

Applications

- Applications similar to HMM
- Labeling or parsing of sequential data
  - Natural language text
  - Biological sequences
    - Classification of proteins
    - Prediction of the secondary structure of DNS and proteine
  - Image recognition and image resauration

Conditional random fields

Discussion

- Generative models
  - HMMs, stochastic grammars, ...
  - Assign joint probability to paired observations
- Discriminative models
  - Maximum entropy Markov models, Conditional random fields, ...

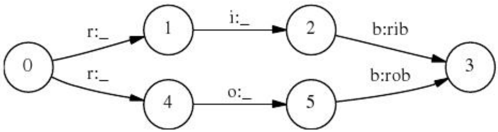
Conditional random fields

Discussion

- Problem of generative models
  - To define joint probability over observation sequences (e.g. words or nucleotides), all possible observation sequences are enumerated
  - Not practical
    - Multiple interacting features
    - Long range dependencies
  - Conditional probability depends on fixed, dependent features
    - Instead of arbitrary independent features
  - Very strict independence assumptions on observations
    - e.g. conditional independence

Conditional random fields

The label bias problem

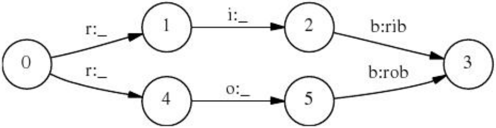


- Problem of (classical) discriminative models (e.g. MEMM)
  - Label bias problem[8]
    - Conservation of score mass
    - States with fewer outgoing transitions are tendentially biased
    - States with low-entropy next state distributions will take little notice of observations

[8] L. Bottou, *Une approche theorique de l'apprentissage connexionniste: Applications a la reconnaissance de la parole*, PhD-thesis, 1991.

Conditional random fields

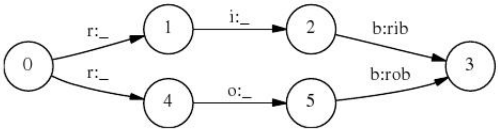
The label bias problem



- Solutions proposed to solve the label bias problem
  - Change state transition structure of the model
    - Collapse states (here: 1 and 4)
    - Delay branching until discriminating observation
  - Start with fully connected model
    - Let training figure out good structure

Conditional random fields

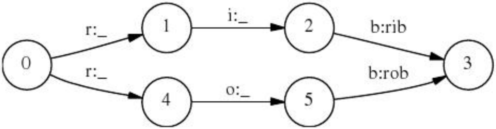
The label bias problem



- Problems with the solutions proposed
  - Change state transition structure of the model
    - Not always possible
    - May lead to combinatorical explosion
  - Start with fully connected model
    - Preludes use of valuable prior structural knowledge

Conditional random fields

The label bias problem



- Requirement for proper solutions
  - Model that accounts for whole state sequences at once
    - Let transitions 'vote' more strongly than others
    - Score mass will not be conserved
    - Transitions can amplify or dampen received mass

Conditional random fields

Discussion

- Conditional random fields
  - Solve label bias problem
  - Single exponential model for joint probability of sequences
  - Less impacted by higher-order dependencies between states

Conditional random fields

Algorithmic model

Definition: Conditional random field

Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$. Then (X, Y) is a conditional random field when the random variables Y_v obey the Markov property with respect to the graph:
$p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)$

- $w \sim v$ means that w and v are neighbours in G
- A CRF is a random field conditioned on the observation X
- In general, the graphical structures of X and Y are not the same.

Conditional random fields

Random field

- Random field
  - Generalization of a stochastic process
  - Underlying parameter need no longer be a simple real
  - Can instead be multidimensional vector space

Conditional random fields

Random field

Random field

Let $S = X_1, ..., X_n$, with the X_i in $\{0, 1, ..., G-1\}$ being a set of random variables on the sample space $\Omega = \{0, 1, ..., G-1\}^n$. A probability measure π is a random field if, for all ω in Ω, $\pi(\omega) > 0$.

Conditional random fields

Example: HMM-like CRF

$$
\begin{aligned}
f_{y',y}(< u, v >, y|_{<u,v>}, x) &= \delta(y_u, y')\delta(y_v, y) \quad (37) \\
g_{y,x}(v, y|_v, x) &= \delta(y_v, y)\delta(x_v, x) \quad (38)
\end{aligned}
$$

- Feature for each state pair (y, y') and each state-observation pair (y, x)

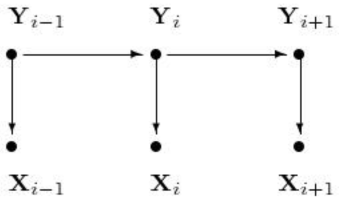
Conditional random fields

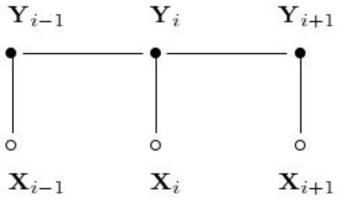
Example: HMM-like CRF

- CRF more expressive: Can model more cases than HMM
- Features do not need to specify completely a state or observation.
  - Therefore, model can be estimated from less training data
- CRFs share all convexity properties of general maximum entropy models

Conditional random fields

Example: HMM-like CRF

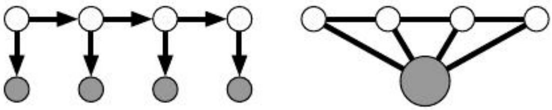




- Graphical structures of HMMs and CRFs
  - Circle indicates that variable is not generated by the model

Conditional random fields

Example: HMM-like CRF



- CRF: Entire observation sequence combined

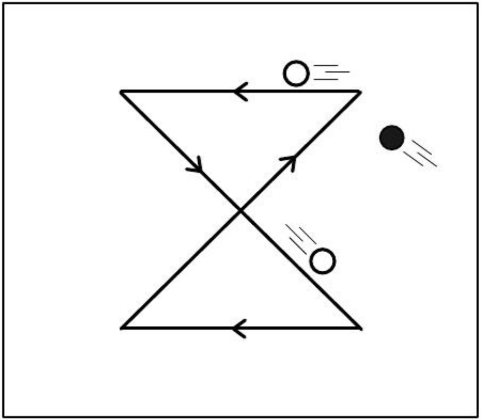
Conditional random fields
Experiments

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM+	4.81%	26.99%
CRF+	4.27%	23.76%

+Using spelling features

Conditional random fields

Experiments



- Seeker robot (black) tries to tag on of the other players
- Simplified variant: Non-seeker robots follow hourclass pattern

Conditional random fields

Experiments

- CRF and HMM accuracy for identifying the seeker

	Hourglass			Unconstrained		
Features	HMM Acc.	CRF Acc.	$\ell(Y\|X)$	HMM Acc.	CRF Acc.	$\ell(Y\|X)$
Positions	33.1	53.6	-959.7	37.1	37.8	-1354.5
Velocities	68.4	89.4	-717.1	55.7	70.4	-1206.5
Velocity Thresholds						
W = $\frac{1}{60}$th sec.	62.5	71.2	-818.0	46.8	58.6	-1148.6
W = 0.1 sec.	63.0	73.9	-784.3	46.0	62.4	-1099.2
W = 0.5 sec.	63.6	80.6	-708.8	68.9	71.9	-983.1
W = 1.0 sec.	60.2	83.1	-721.8	67.8	75.3	-980.9
W = 1.5 sec.	56.9	85.5	-731.7	68.8	77.8	-1004.7
W = 2.0 sec.	53.7	87.1	-751.1	72.1	77.3	-1036.3
Chasing	75.8	95.4	-622.3	65.5	80.4	-1058.3
Distance (U)	46.6	49.5	-779.7	43.5	42.3	-604.4
Distance (N)	46.6	49.9	-200.6	43.5	58.0	-223.4
Distance & Chasing (U)	75.6	99.3	-90.6	65.8	93.9	-181.8
Distance & Chasing (N)	75.6	99.3	-115.3	65.8	97.6	-112.2
Many Features	72.4	98.1	-172.2	63.4	98.5	-178.9
Redundant Features	72.4	95.7	-509.3	52.7	93.8	-6432.3

Context prediction with CRF

Prediction procedure

- Context prediction with CRF:
  1. Given the transition model, estimate all probabilites between states and state action probabilites
  2. Given a sequence V^T, decode the most probable sequence of hidden states
  3. Extrapolate the sequence of expected hidden states

Aspects of prediction algorithms
Summary

	IPAM	ONISI	Markov	CRF
Numeric Contexts	yes	no	no	no
Non-numeric Contexts	yes	yes	yes	yes
Complexity	$O(k)$	( )	$O(C^2)$	$O(C^2)$
Learning ability	(no)	yes	yes	yes
Approximate matching	no	no	no	no
Multi-dim. TS	(no)	(no)	(no)	(no)
Discrete data	yes	yes	yes	yes
Variable length patterns	no	yes	no	(yes)
Multi-type TS	yes	no	(no)	(no)
Continuous data	no	no	no	no
Pre-processing	$O(k)$	–	$O(k)$	$O(k)$
Context durations	no	no	no	no
Continuous time	no	no	yes	yes

Properties of CRF prediction approaches

Conclusion

- CRF processes are straightforward and easily applied to context prediction tasks.
- Model can be applied to numerical and non-numerical data alike.
- Prediction that reaches farther into future implicitly utilises already predicted data which might consequently decrease the prediction accuracy.

Conclusion